A Voice-Activated Web-based Mandarin Chinese Spoken Document Retrieval System

Hsin-min Wang, Berlin Chen, Liang-jui Shen, and Chao-chi Chang

Institute of Information Science, Academia Sinica
128 Academia Road, Section 2, Taipei 115, Taiwan, Republic of China
Email: whm@iis.sinica.eud.tw

Abstract

This paper presents a working voice-activated web-based Mandarin Chinese spoken document retrieval system. This system has integrated technologies of both spoken document retrieval and voice-activated WWW browser. The target database to be retrieved consists of tens of hours of radio and television Mandarin Chinese broadcast news. Extensive experiments have been conducted and a prototype system has been successfully implemented.

Keywords: spoken document retrieval; speech recognition; speech interface; voice-activated browser; WWW; Mandarin Chinese.

1 Introduction

Massive quantities of audio and multimedia content, such as broadcast radio and television programs, are becoming available in the global information infrastructure. As a result, spoken document retrieval (SDR) has been extensively studied in recent years [1-3]. In addition, a speech interface that allows easy access to information on the WWW has the potential to make the browser more friendly and powerful [4-6]. Accordingly, in this paper, we integrate technologies of spoken document retrieval and voice-activated browser to build a voice-activated web-based Mandarin Chinese spoken document retrieval system.

Recently, spoken document retrieval applications are crossing the threshold of practicality, as evidenced by Compaq's SpeechBot [7], which is a web-based English spoken document retrieval system, where the queries are textual, and documents are in audio form. Using spoken queries to access the audio streams remains a very challenging research topic because the query terms could contain recognition errors as well and such errors could make the retrieval system completely mistake what the user needs especially when the query is short. There are still not many reports on this very challenging task. However, since wireless communication is becoming increasingly popular, development of technology for voice retrieval of speech information is becoming increasingly important.

The rest of this paper is organized as follows. Section 2 gives a brief introduction to Mandarin Chinese speech recognition. Section 3 presents our spoken document retrieval approach. The voice-activated browser is introduced in Section 4 while the voice-activated web-based Mandarin Chinese spoken document retrieval system is presented in Section 5. Finally, The concluding remarks are given in Section 6.

2 Speech Recognition

2.1 A Brief Introduction to Mandarin Chinese

In the Chinese language, there are more than 10,000 characters. A Chinese word is composed of from one to several characters. The combination of one to several of such characters gives an almost unlimited number of words, which can be found in different versions of dictionaries and texts on different subjects. Two nice features of the language are that all characters are monosyllabic, and that the total number of phonologically allowed syllables is only 1,345. Furthermore, Mandarin Chinese is a tonal language, in which each syllable is assigned a tone and there are a total of 4 lexical tones plus 1 neutral tone. If the differences among the syllables caused by tones are disregarded, only 416 "base syllables" (i.e., the tone-independent syllable structures) instead of 1,345 different "tonal syllables" (i.e., including the tone features) are required to cover all the pronunciations for Mandarin Chinese [8]. Base syllable recognition is, thus, believed to be the first key problem in Mandarin Chinese speech recognition as well as in spoken document retrieval.

2.2 Signal Processing

For speech recognition, typically, a speech signal is divided into a number of overlapping time frames, and a speech parametric vector is computed to represent each time frame. In our speech recognizer, spectral analysis is applied to a 20 msec frame of speech waveform every 10 msec. For each speech frame, 12 mel-frequency cepstral coefficients (MFCC) and the logarithmic energy are extracted, and these 13 coefficients along with their first and second time derivatives are combined together to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to all the training sentences, spoken documents and speech queries.

2.3 Acoustic Modeling

Although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of the Chinese language, using it leads to inefficient utilization of the training data in the training phase and high computation requirements in the recognition phase. Thus, the acoustic units chosen here are 112 context-dependent INITIAL's and 38 context-independent FINAL's based on the monosyllabic nature of Mandarin Chinese and the INITIAL-FINAL structure of Mandarin base syllables. Here, INITIAL is the initial consonant of the base syllable, and FINAL is the vowel (or diphthong) part but including optional medial or nasal ending.

Each INITIAL is represented by a HMM with 3 states while each FINAL is represented by one with 4 states. The Gaussian mixture number per state ranges from 2 to 16, depending on the amount of training data. Therefore, every syllable unit is represented by a 7-state HMM. The silence model is a 1-state HMM with 32 Gaussian mixtures trained by using the non-speech segments.

2.4 Syllable Recognition

A multiple-pass search strategy is used in the syllable recognition. In the first pass, the Viterbi search is performed based on the acoustic models and the syllable bigram language model, and the score at every time index is stored. In the second pass, a backward time-asynchronous A* tree search [9] generates the best syllable sequence based on the heuristic scores obtained from the first pass search and the syllable trigram language model. In the third pass, based on the state likelihood scores calculated in the first pass and the syllable boundaries of the best syllable sequence, the syllable recognizer further performs the Viterbi search on each utterance segment which may include a syllable and outputs several most likely syllable candidates, and a syllable lattice can thus be constructed. In the fourth pass, the word graph is constructed from the syllable lattice first and, then, a Viterbi search is conducted on the word graph to find the best word sequence using the word unigram and bigram language models. Finally, the word sequence is converted into the syllables by pronunciation lookup using a Mandarin Chinese pronunciation lexicon.

2.5 Keyword Spotting

The search framework of keyword spotting is based primarily on a lexical network concatenated with a left filler model and a right filler model as shown in Figure 1. Each arc of the lexical network represents a subword unit, thus the network is able to handle any arbitrarily assigned vocabulary set. Here, the syllable is chosen as the subword unit and each syllable is composed of two sub-syllabic models, as mentioned in Section 2.3. 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 cover the surrounding non-keyword part of the input utterances. Our keyword spotting process is based on

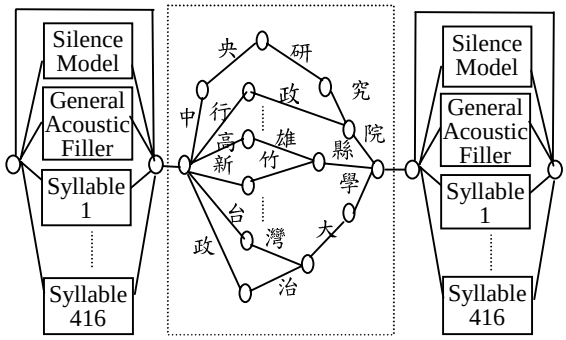

Figure 1: The search framework of keyword spotting.

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 score stored at every time index. At the same time, a compact syllable lattice with the respective lattice node scores evaluated left-to-right including the left filler model scores is generated from the lexical network based on the constraint grammar considering the keyword structure. 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 first pass. In the A* search process, in every iteration a partial path with the maximal evaluation function is popped, extended, and stored in a stack, and the stack is sorted by the evaluation functions of all the extended partial paths. This process is terminated when the N-best complete paths (candidate keywords) are obtained sequentially.

Syllable Segments	Examples
S(N), N=1	$(s_1)\ (s_2)\ \ldots(s_{10})$
S(N), N=2	$(s_1\ s_2)\ (s_2\ s_3)\ldots(s_9\ s_{10})$
S(N), N=3	$(s_1\ s_2\ s_3)\ (s_2\ s_3\ s_4)\ldots(s_8\ s_9\ s_{10})$
Syllable Pair Separated by *n* Syllables	Examples
P(n), n=1	$(s_1\ s_3)\ (s_2\ s_4)\ \ldots(s_8\ s_{10})$
P(n), n=2	$(s_1\ s_4)\ (s_2\ s_5)\ \ldots(s_7\ s_{10})$
P(n), n=3	$(s_1\ s_5)\ (s_2\ s_6)\ \ldots(s_6\ s_{10})$

Table 1: The indexing terms extracted from an example syllable string $S_1\ S_2\ S_3\ldots\ S_{10}$.

3 Spoken Document Retrieval

This section will introduce our syllable-based Mandarin Chinese spoken document retrieval approach.

3.1 Syllable-level Indexing Terms

In our syllable-based Mandarin Chinese spoken document retrieval approach, the syllable-level indexing terms composed of the overlapping syllable segments with length *N* (*S(N), N=1~3*) and the syllable pairs separated by *n* syllables (*P(n), n=1~3*). Considering a syllable string of *10* syllables $S_1\ S_2\ S_3\ldots\ S_{10}$, examples of the former are listed

on the upper half of Table 1, while examples of the latter on the lower half of Table 1. The overlapping syllable segments with length *N* can capture the information of polysyllabic words or phrases while the syllable pairs separated by *n* syllables can tackle the problems arising from the flexible wording structure, abbreviation, and speech recognition errors.

3.2 Information Retrieval Model

Vector space models widely used in many text information retrieval systems were used here. A document is represented as a set of feature vectors, each consists of one type of indexing terms (*S(1), S(2), S(3) and P(1~3)*). Each component of a feature vector $\vec{v}$ is associated with the weighted statistics $ws_{\vec{v}}(t)$ of a specific indexing term *t*:

$$ws_{\vec{v}}(t) = (1 + \ln(\sum_{i=1}^{n_t} cm_t(i))) \cdot \ln(N / N_t), \quad (1)$$

where $cm_t(i)$, ranging from *0* to *1*, is the normalized acoustic confidence measure of the *i*-th occurrence of indexing term *t* within the document, and n_t is the total occurrences of indexing term *t* in the document. The value of $1 + \ln(\sum_{i=1}^{n_t} cm_t(i))$ denotes the term frequency of indexing term *t* and the logarithmic operation is used to condense the distribution of the term frequency. $\ln(N/N_t)$ is the Inverse Document Frequency (IDF), where *N* is the total number of documents in the collection and N_t is the number of documents that contain term *t*. A query is also represented by a set of feature vectors based on the same feature vector construction procedure except that $cm_t(i)$ in Equation (1) is simply set to *1* for all *t* and *i* when a text query is given; i.e., using a perfect manual transcription. The Cosine measure is used to estimate the query-document similarity for each type, *s*, of indexing terms:

$$SIM_s(\vec{q}_s, \vec{d}_s) = \frac{(\vec{q}_s \bullet \vec{d}_s)}{(\|\vec{q}_s\| \cdot \|\vec{d}_s\|)}. \quad (2)$$

The overall similarity is then the weighted sum of the similarity scores of all types of indexing terms:

$$SIM(\vec{q}, \vec{d}) = \sum_s w_s \cdot SIM_s(\vec{q}_s, \vec{d}_s), \quad (3)$$

where w_s are tunable weights. Currently, they are 0.1, 0.7, 0.3 and 0.5 for *S(1), S(2), S(3)* and *P(1~3)*, respectively.

3.3 Experimental Results

The target database to be retrieved in the following experiments consists of Mandarin Chinese broadcast news collected from December 1998 to July 1999. There are 757 recordings (about 10.2 hours of speech materials) collected from Broadcasting Corporation of China (BCC). Each recording is a short news abstract (~50 seconds) produced by an anchor speaker, and contains several news items. Some recordings contain background music. 40 short queries were selected to support the retrieval experiments. Each query contains roughly only 4 characters (or syllables) and has on average 23.3 relevant documents among the 757 in the database, with the exact number ranging from 1 to 75. Two (one male and one female) speakers pronounced the 40 query terms, respectively.

Another Mandarin Chinese broadcast news database consists of 453 stories (about 6.0 hours of speech materials) collected from Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT), all located at Taipei, was used for training the acoustic models for recognition of the spoken documents. All recordings were manually transcribed and segmented into stories and sentences. In addition, a Mandarin Chinese read speech database with 5.3 hours of speech materials for phonetically balanced sentences and isolated words produced by roughly 120 speakers was used for training the acoustic models for recognition of the speech queries. The syllable-based and word-based *N*-gram language models were trained by a newswire text corpus consisting of 80 million Chinese characters collected from Central News Agency (CNA) in 1999. Notice that both the newswire text corpus and the broadcast news database were collected almost in the same time frame. Word segmentation and phonetic labeling were performed for the training materials using a 62k-word Mandarin Chinese lexicon.

The syllable accuracies for the spoken documents and the speech queries are 73.37% and 89.20%, respectively. The retrieval performance for using speech queries and text queries in terms of *non-interpolated average precision* [10] is 0.674 and 0.715, respectively.

4 The Voice-Activated Browser

Figure 2 depicts the interaction between the speech interface, browser, and the WWW. To allow any user to use the voice-activated browser naturally without training, 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, as mentioned in Section 2.3, to allow the recognition of any vocabulary word without training on that specific word.

All the commands are automatically defined as keywords and are added to the dynamic keyword lexicon. For bookmarks, user-defined key phrases are also defined as keywords and, thus, are 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 is first applied to obtain the pronunciation according to a general Chinese lexicon, 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. If the full name is less than 11 characters, the full name itself is also thought of as a long keyword. While for the new bookmarks and visible link names, exactly the same word identification and keyword extraction procedure can be applied. Note that the word identification and keyword extraction procedure must be 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

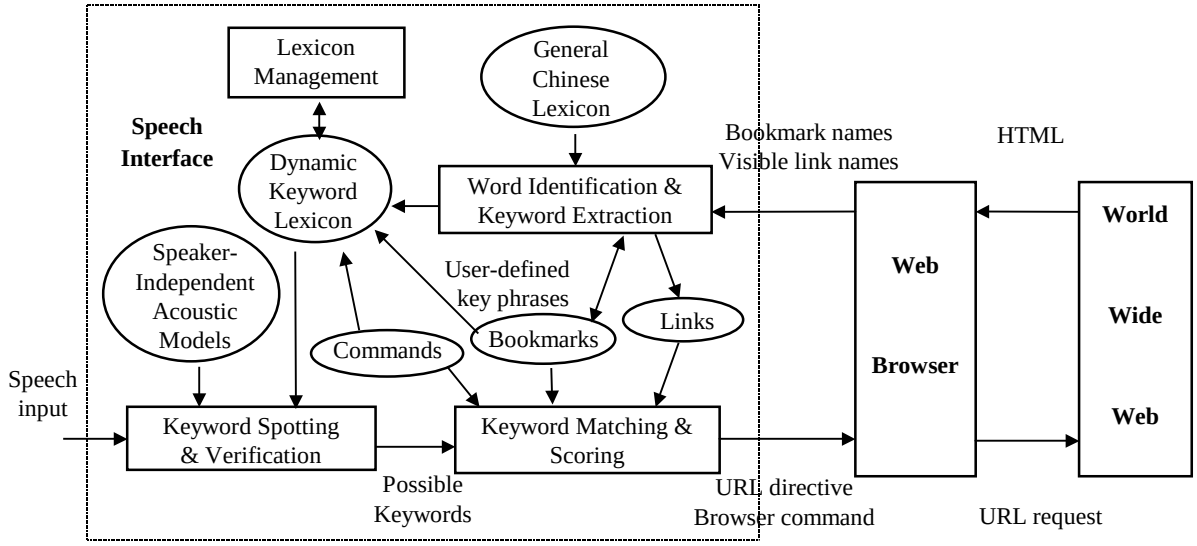

Figure 2: Interaction between the speech interface, browser, and the WWW.

high keyword spotting accuracy, the size of the dynamic keyword lexicon must be limited to a pre-defined number. The lexicon management module is, thus, designed to keep the lexicon size under a pre-defined number based on the first-in-first-out criterion. Currently, 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 on many previously visited pages. The users can go to any link on the previously visited pages without exhaustedly clicking on the back bottom to find that link.

Given a speech input, the keyword spotter first generates a set of possible keyword candidates. Then, the utterance verification mechanism, which acts as a post-processor, further verifies these keyword candidates. 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. The speech input can be any combination of words or characters that comprise the corresponding bookmark or link, and the users do not need to input the full name of a link or a bookmark. The interface also allows users to input speech like "go to *link* (or *bookmark*) *name*", "I want …", etc.

4.1 Keyword Matching and Scoring

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

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

where S_i is the similarity score of the i-th sentence (each command, bookmark, or link is thought of 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 the keywords that are matched from the spotted keyword candidates and, at the same time, these keywords are with the higher confidence measures will have the higher priority to be chosen as the output. The following three weights are used to improve the accuracy of the keyword matching and scoring procedure. 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 matched keyword j while D is the duration of the speech utterance. The first weight makes the sentence with a higher percentage of the matched characters have the higher priority to be chosen, while the second weight makes the sentence with a higher percentage of the matched duration have the higher priority to be chosen. The basic idea for the third weight is that the sentence with a longer average length of the matched keywords should have the higher priority to be chosen because a long keyword is usually more reliable than a short one. Thus, $f(x)$ can be any monotonically increasing function. Here we use a simple linear function with $f(10)=1$ and $f(1)=0.5$.

4.2 Experimental Results

The following experiments are based on 104 Chinese web pages. 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 acoustic models are the same as that used in Section 3.3 for the recognition of speech queries. The experimental results

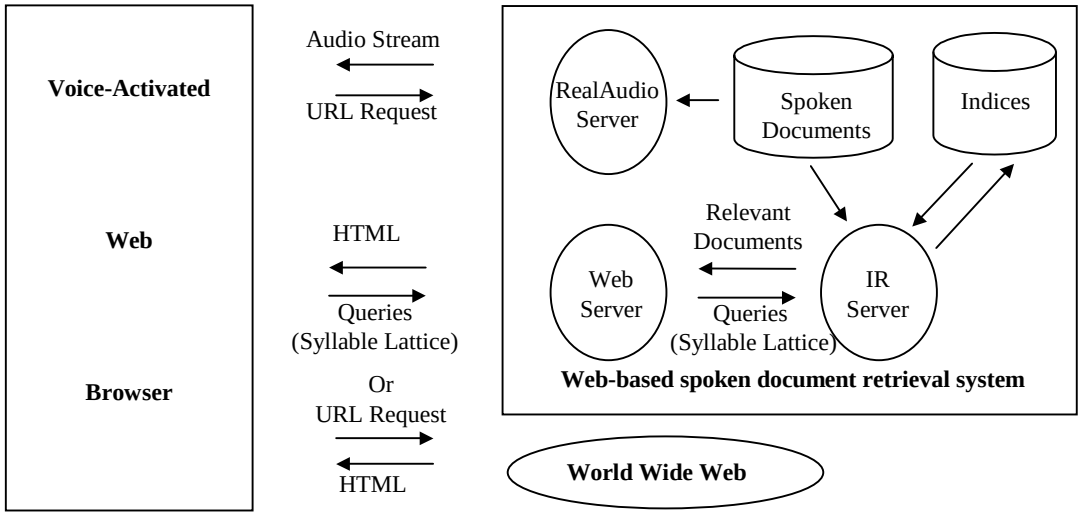

Figure 3: Interaction between the voice-activated browser, the WWW and the web-based spoken document retrieval system.

are summarized in Table 2. It can be found that very high accuracy can be achieved if the input utterances contain only pure control commands; i.e., the Cmd case, or full link names without any irrelevant word; i.e, the FullLk case. 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, browsing the web pages using quasi-natural-language speech is the most natural and attractive way and is highly desired. Better approaches to further improve the performance of such cases are very important.

	Male		Female	
	# Utterance	Accuracy	# Utterance	Accuracy
Cmd	183	0.9289	130	0.9538
FullLk	304	0.9638	306	0.9215
CmdIrWd	114	0.8070	121	0.8016
FullLkIrWd	134	0.8283	166	0.9156
KeyLk	115	0.8956	189	0.8095
Average	-	0.9047	-	0.8859

Table 2: The testing results for the voice-activated browser.

5 The Voice-Activated Web-Based Mandarin Chinese Spoken Document Retrieval System

5.1 The Prototype System

We have integrated the voice-activated browser and the spoken document retrieval approach into a voice-activated web-based spoken document retrieval system. Figure 3 depicts the interaction between the voice-activated browser, the WWW and the web-based spoken document retrieval system. This system allows users to browse the Chinese web pages using unconstrained Mandarin speech in a variety of ways as described in Section 4. This function is depicted in the lower half part of Figure 3. In addition, the users can input "*Search for xxx from the video database*" to retrieve the desired spoken documents from the spoken document database. First of all the voice-activated browser will send the syllable lattice recognized from the speech query to the web server of the spoken document retrieval system, as shown in the upper half part of Figure 3. Then, the web server will pass the syllable lattice to the information retrieval (IR) server and the IR server will return a list of relevant documents to the web server, through the retrieval process as described in Section 3. Finally, the web server will return a HTML file, which contains a list of relevant documents and the URLs of all the relevant documents. The users can play the audio or video document by speaking the name of the hyper-link or by clicking on the hyper-link using mouse. Of course, this interface also provides the text-input mode for users to use text queries.

5.2 System Performance

We have conducted some preliminary retrieval experiments on the prototype system. The spoken document database consists of 53.3 hours (2619 documents) of Mandarin Chinese broadcast news, 10.2 hours (757 documents) of them are from what we used in evaluating the spoken document retrieval approach in Section 3.3. They are collected on air, and all speech is from anchor speakers. The syllable recognition accuracy for this set is 73.37%. The other 43.1 hours (1862 documents) are in RealAudio format originally. They contain a large amount of speech from field reporters and interviewees, and are with very low resolution due to severe compression. These characteristics inevitably cause difficulty in speech recognition. We have randomly selected 29 stories (50.7 minutes) for measuring speech recognition performance. The syllable recognition accuracies for the anchor (8.2 minutes) and non-anchor (42.5 minutes) speech are 37.08%

and 25.86%, respectively, and the average accuracy is only 27.87%.

Currently, we have only tested the prototype spoken document retrieval system using text queries. We use the same query set (40 short queries) as in Section 3.3. For each query, the system returned 20 documents. Since the relevance judgment for the new database is not available, we are not able to obtain the traditional recall/precision graph or non-interpolated average precision, which must be measured at standard recall levels [1]. Hence, the first evaluation measure used here is an average of the precision for each query after a given number of documents have been retrieved for that query. The retrieval performance is plotted in Figure 4. The average precisions are 0.90, 0.815, 0.715, 0.627, and 0.55, respectively, when 1, 5, 10, 15, and 20 documents were retrieved. It was found that the performance degraded fast when the number of retrieved documents increased. This is obviously because some queries only have few relevant documents; i.e., for these queries, increasing the number of retrieved documents simply includes more irrelevant documents, which hurt precision. For example, we found that for some queries only the top documents are relevant while all the rest are irrelevant. As a result, the second evaluation measure used is a modified non-interpolated average precision. The precision average for each query is obtained by computing the precision after every retrieved relevant document and then averaging these precisions over the total number of retrieved relevant documents, but now the evaluation is based on the top N documents retrieved instead of the complete rank list. The average precision obtained in this manner will be higher than the ideal non-interpolated average precision and the average precision used in the first evaluation measure. Using the second evaluation measure, the average precisions are 0.896, 0.882, 0.872, and 0.865, respectively, when 5, 10, 15, and 20 documents were retrieved. The high average precisions indicate that the retrieved relevant documents tend to gather in the top group of documents retrieved. Furthermore, it's worth mentioning that 90% of the queries are able to get the relevant document when only one document is returned while 95% of queries are able to get at least one relevant documents within 3 retrieved documents. By taking advantage of this situation, we can apply relevance feedback techniques [3] to enhance retrieval performance.

6 Conclusions

We have described a working speech interface, which enables users to browse the Chinese web pages and retrieve audio streams such as broadcast news using unconstrained

[1] The precision average for each query is obtained by computing the precision after every retrieved relevant document and then averaging these precisions over the total number of retrieved relevant documents. These query averages are then averaged across all queries to create the non-interpolated average precision. [10]

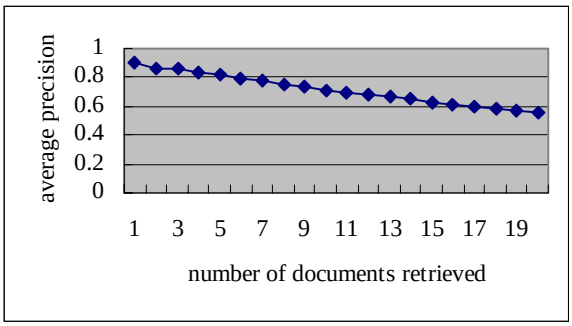

Figure 4: The retrieval performance of the prototype system (the average precision with respect to the number of documents retrieved).

Mandarin speech in a variety of ways. Because wireless communication is becoming very popular, development of technologies for browsing the web pages and retrieving the multimedia information via voice is becoming increasingly important. Our experimental results have indicated the potential in this direction, though there are still many unsolved problems, such as the speech recognition on the mobile phone.

References

1 K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, "Experiments on Spoken Document Retrieval," *Information Processing & Management*, 32(4), pp. 399-417, 1996.

2 M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition," Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

3 B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of Mandarin Broadcast News using Spoken Queries," *ICSLP2000*.

4 C. T. Hemphill and P. R. Thrift, "Surfing the Web by Voice," *Multimedia95*, pp. 215-222.

5 K. Kondo and C. T. Hemphill, "Surfing the World Wide Web with Japanese," *ICASSP97*, pp. 1151-1154.

6 H. M. Wang, Y. H. Chou, and B. Chen, "Browsing the Chinese Web Pages Using Mandarin Speech," *International Journal of Computer Processing of Oriental Languages*, 13(1), pp. 35-51, March 2000.

7 B. Logan, P. Moreno, J. M. Van Thong, and E. Whittaker, "An Experimental Study of An Audio Indexing System for the Web," *ICSLP2000*.

8 L. S. Lee, "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 14(4), pp. 63-101, 1997.

9 P. Kenny, R. Hollan, V. N. Gupta, M. Lennig, P. Mermrlstein, and D. O'Shaughnessy, "A*-Admissible Heuristics for Rapid Lexical Access," *IEEE Tran. on ASSP*, 1(1), January 1993.

10 D. Harman, "Overview of the Second Text REtrieval Conference (TREC-2)," TREC-2, pp. 1-20.