LARGE-VOCABULARY CHINESE TEXT/SPEECH INFORMATION RETRIEVAL USING MANDARIN SPEECH QUERIES

Bo-ren Bai[1], *Berlin Chen*[2], *Hsin-min Wang*[2], *Lee-feng Chien*[2], *and Lin-shan Lee*[1,2]
[1]Department of Electrical Engineering, National Taiwan University, Taipei
[2]Institute of Information Science, Academia Sinica, Taipei
E-mail: white@speech.ee.ntu.edu.tw {berlin, whm, lfchien, lsl}@iis.sinica.edu.tw

ABSTRACT

The network technology and the Internet are creating a completely new information era. It is believed that in the near future numerous of digital libraries and a great variety of multimedia databases, which consist of heterogeneous types of information including text, audio, image, video and so on, will be available worldwide via the Internet. This paper deals with the problem of Chinese text and Mandarin speech information retrieval with Mandarin speech queries. Instead of using the syllable-based information alone, the word-based information was also successfully incorporated to further improve the retrieving performance. A prototype system with an interface supporting some user-friendly functions was successfully implemented and the initial test results verified the feasibility of our approaches.

1. INTRODUCTION

The network technology and the Internet are creating a completely new information era. It is believed that in the near future, numerous of digital libraries and a great variety of multimedia databases will be available worldwide via the Internet. The digital libraries and multimedia databases will consist of heterogeneous types of information including text, audio, image, video and so on. Intelligent and efficient information retrieval techniques allowing easy access to huge amount and various types of information become highly desired and have been extensively studied in recent years [1-2]. This paper deals with the problem of Chinese text and Mandarin speech information retrieval with Mandarin speech queries. To achieve this purpose, key technologies including Chinese text keyword extraction [3] and Mandarin speech keyword spotting [4] were first developed, then based on these key technologies new approaches were developed integrating relevant technologies including text processing, speech recognition, and information retrieval.

Sub-word-based features have been found useful in retrieval using speech queries since they can somehow handle the out-of-vocabulary problem to a certain degree [2], but in any case they carry significantly less semantic information, thus limit the potential for more precise retrieval. Considering the characteristic monosyllabic structure of the Chinese language, syllable-based approach has been found to be an attractive special case for the sub-word-based approaches for retrieving Chinese text [5] and speech [6] information using speech queries. On the other hand, word-based features, specially the keywords, carry plenty of precise semantic information for retrieval. However, it is inevitable that some important word (or keyword) information will be lost when either the queries or the database records are in form of speech due to speech recognition errors. Also, for large-vocabulary information retrieval the out-of-vocabulary problem always exists regardless of the size of the lexicon used, i.e., some important information-carrying words (or keywords) are out of the used lexicon. Recently, very efficient techniques to extract keywords from text documents have been successfully developed for retrieving Chinese text information [3]. Furthermore, even if the desired information is in form of speech, it is always possible for the user to have in hand a text collection with subject domain relevant to the desired topic. Therefore the keywords automatically extracted from the text collection is always very helpful in retrieval, regardless of whether the target database records are in form of text or speech. Thus in this paper we integrate the advantages of the syllable and word (or keyword) information properly for retrieval of large-vocabulary Chinese text and speech information using Mandarin speech queries. A prototype system with an interface supporting some user-friendly functions was successfully implemented and the initial test results verified the feasibility of the technologies and approaches.

The rest of this paper is organized as follows. The overall architecture of our approach is presented in section 2, and some experimental results are discussed in section 3. Then, the prototype system is introduced in section 4. And, finally, the concluding remarks are made in section 5.

2. OVERALL ARCHITECTURE

The overall architecture of the proposed Chinese text and speech information retrieval system is shown in Figure 1. In this approach, the keyword extraction techniques [3] are first performed to extract automatically an adequate keyword set for all dynamic text records obtained from the Internet. During the retrieving processes, syllable-based feature vectors, including such information as frequency counts and inverse document frequency (IDF) parameters for syllable pairs and single syllables, constructed for both the input speech query and the speech/text records in the databases, are first used for similarity measure to retrieve the top N possible records relevant to the query. At the same time, speech keyword spotting is performed on the input speech query based on the keyword set extracted from the text collection on the

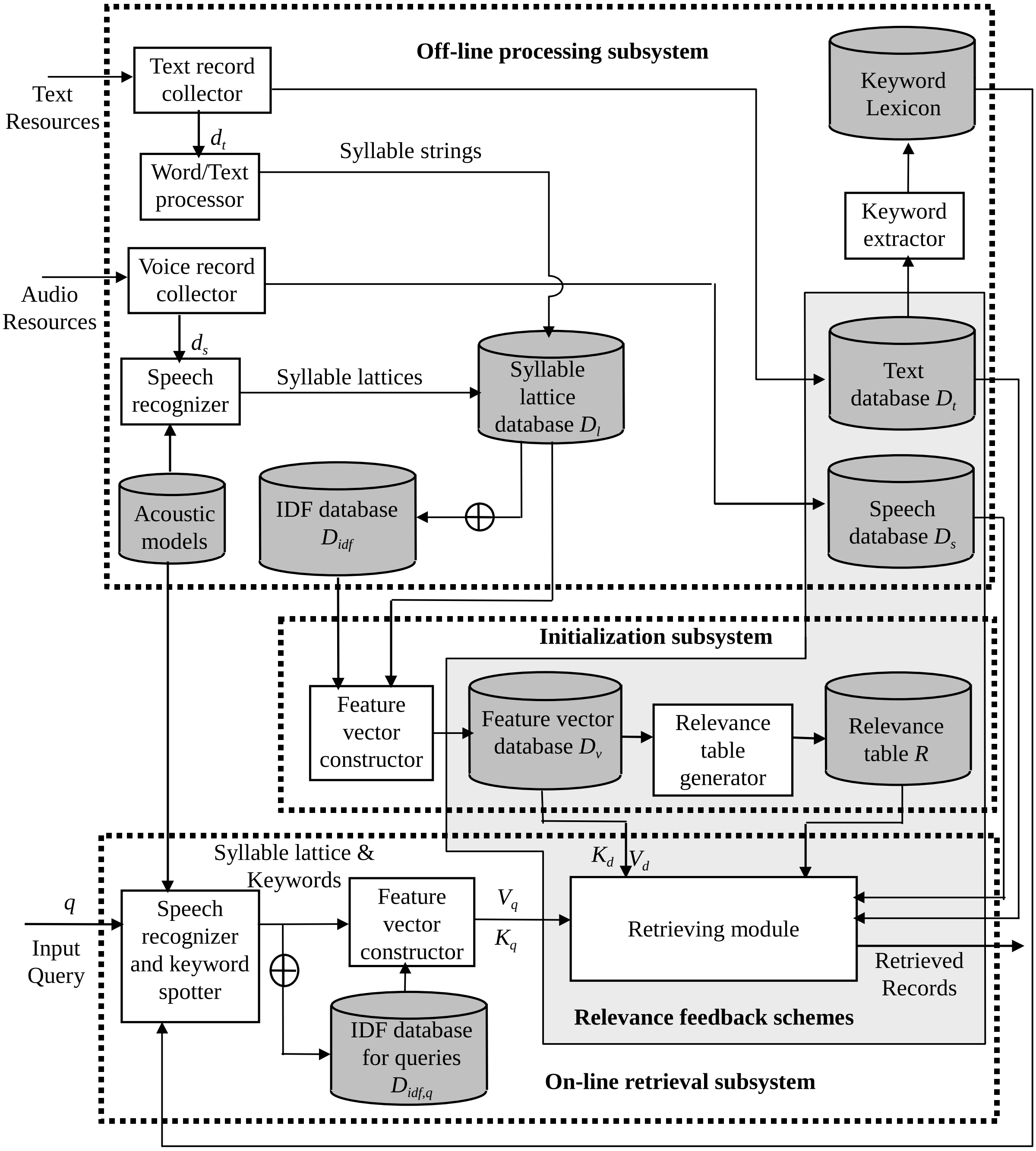


Figure 1: The overall architecture of the proposed Chinese text and speech information retrieval system.

relevant subject domain. For retrieving voice records, these keywords spotted from the input speech query are taken as a new keyword set, and speech keyword spotting process is then performed for those initially selected top *N* relevant voice records. For retrieving text records, text keyword search is performed instead. In both cases the scores for the keywords found in the records are weighted and integrated into the syllable-based scores, so that the initially selected top *N* relevant records are re-ranked to give the final results. Since the speech keyword spotting or text keyword search are performed only for a very small set of keywords and a very small set of initially selected records, only limited extra computation will be needed. In addition, techniques such as relevance feedback were also applied to support interactive retrieval.

3. EXPERIMENTS

The database used for simulation experiments consists of 5,819 text records and 500 voice records on similar subject domains, i.e., they both contain news published in Taiwan although in different time spans. 6,082 keywords extracted from the text collection are used here

	Exact Matching		Near Matching	
	Number of keyword inside lexicon	Coverage percentage (%)	Number of keyword inside lexicon	Coverage percentage (%)
Automatically extracted keyword lexicon	496	47.06	677	64.23
General-purposed word lexicon	293	27.80	541	51.33

Table 1: The numbers and percentages of the keywords manually selected from the testing records covered by the automatically extracted keyword lexicon and the general-purposed word lexicon.

as the keyword lexicon for detection keyword information from both the speech information and the speech queries. To assess the feasibility of using the extracted keyword lexicon to help retrieving the 500 voice records, we will first examine the information coverage of the testing database by the extracted keyword lexicon. First of all, by carefully examining the text materials of the 500 voice records for retrieval, 1,054 keywords are manually chosen to be key information for the testing records. The numbers and percentages of the 1,054 manually selected keywords covered by the automatically extracted keyword lexicon of 6,082 keywords are shown in Table 1, and those of the 1,054 keywords covered by a general-purposed Chinese lexicon of roughly 85,000 words [7] are also shown for comparison. It can be found that with exact matching criterion, i.e., the manually extracted keywords have to be exactly the same as the words in the lexicons, only 47.06% and 27.80% of manually selected keywords in the testing database are covered by the automatically extracted lexicon and the general-purposed lexicon, respectively. While with near matching criterion, i.e., the case that the manually extracted keywords are compound words composed of a few words in the lexicon, is considered correct, 64.23% and 51.33% of keywords in the testing database are covered by the automatically extracted lexicon and the general-purposed lexicon, respectively. It can be found that in both cases the automatically extracted keyword lexicon covers more keywords in the testing records than the general-purposed lexicon.

With the automatically extracted keyword lexicon, the approach to integrate syllable and keyword information for speech information retrieval was evaluated. The result is shown in Figure 2. It can be found that with the keyword information incorporated, the performance can be improved slightly. The average precision rate in this case was improved from 0.4486 to 0.4593. A possible reason for the limited improvement may be that the keyword set of 6,082 is much larger than, and probably quite different from, the real set of keywords specifying the small testing set of 500 voice records. Also, such a large keyword set can degrade the keyword spotting accuracy. Further studies will be needed in this area.

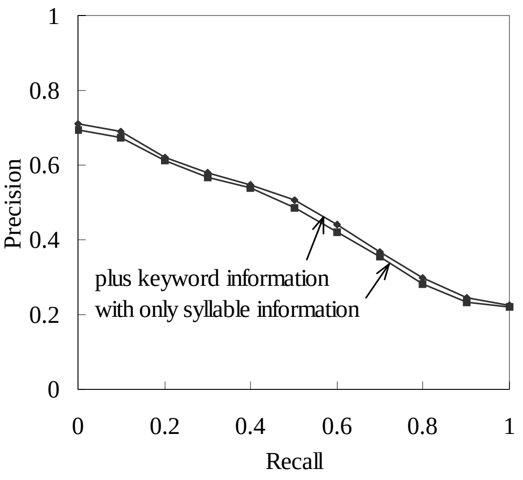
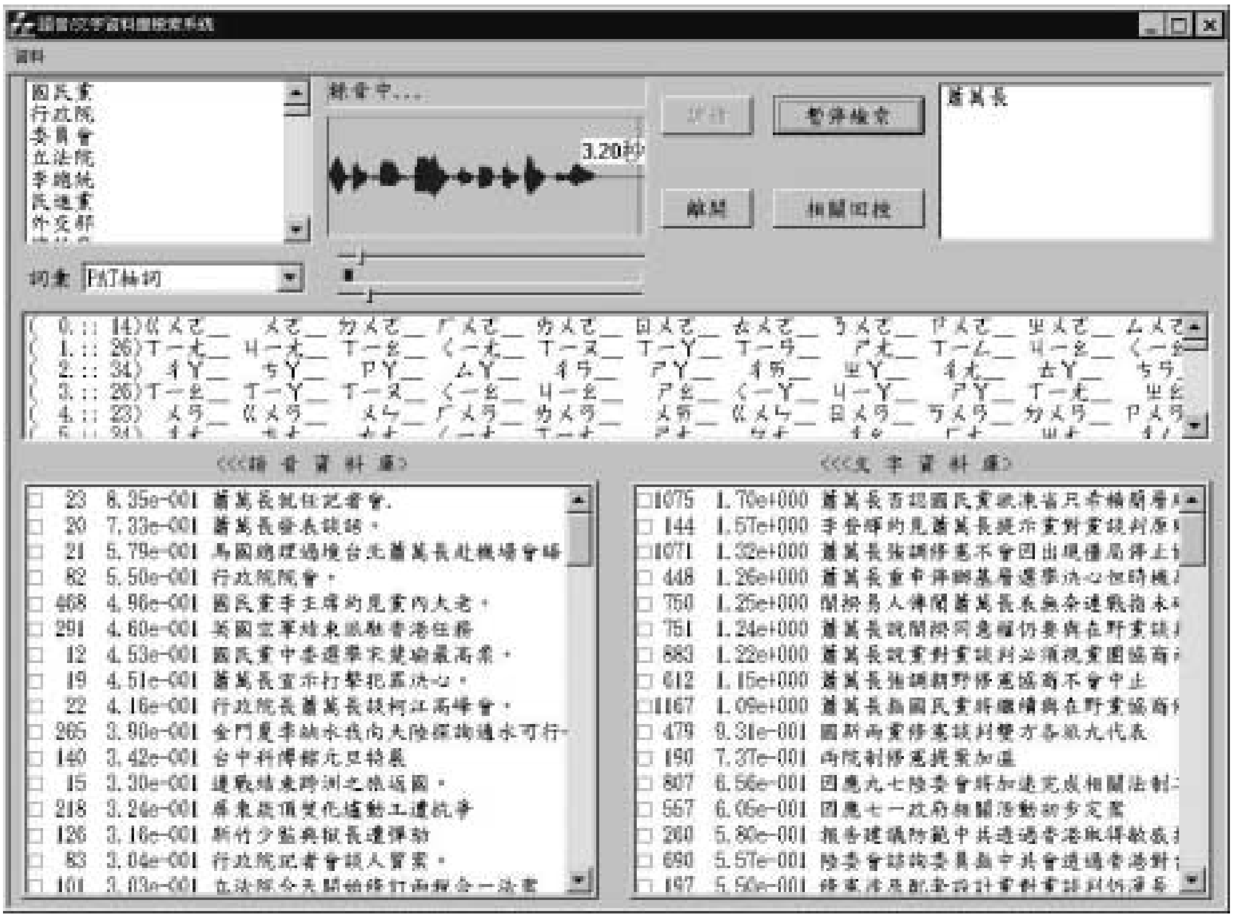


Figure 2: The performance comparison between using syllable information only and using both syllable information and keyword information for speech information retrieval.

4. THE PROTOTYPE SYSTEM

All the above approaches were finally integrated into a prototype system, as shown in Figure 3, for text and speech information retrieval using speech queries. The system includes an interface supporting some user-friendly functions. The upper left sub-window lists the keyword vocabulary used in the system, extracted off-line in advance from the text database. The sub-window right to the keyword list shows the speech waveform of the input query. Then several buttons including “離開(exit)”, “調適(adaptation”), “開始檢索(begin to retrieve)” and “相關回授(relevance feedback)” are designed for corresponding functions. The upper right sub-window shows the keywords spotted from the input speech query by the speech keyword spotter. The middle sub-window below the above upper sub-windows shows the syllable lattice constructed by the continuous speech recognizer. Both the recognized keywords and the syllable lattice will be used for retrieving text and speech databases. The lower left sub-window and lower right sub-window show the retrieved results of the speech database (語音資料庫) and text database (文字資料庫),



```
國民黨
行政院
委員會
立法院
李總統
民進黨
外交部
```

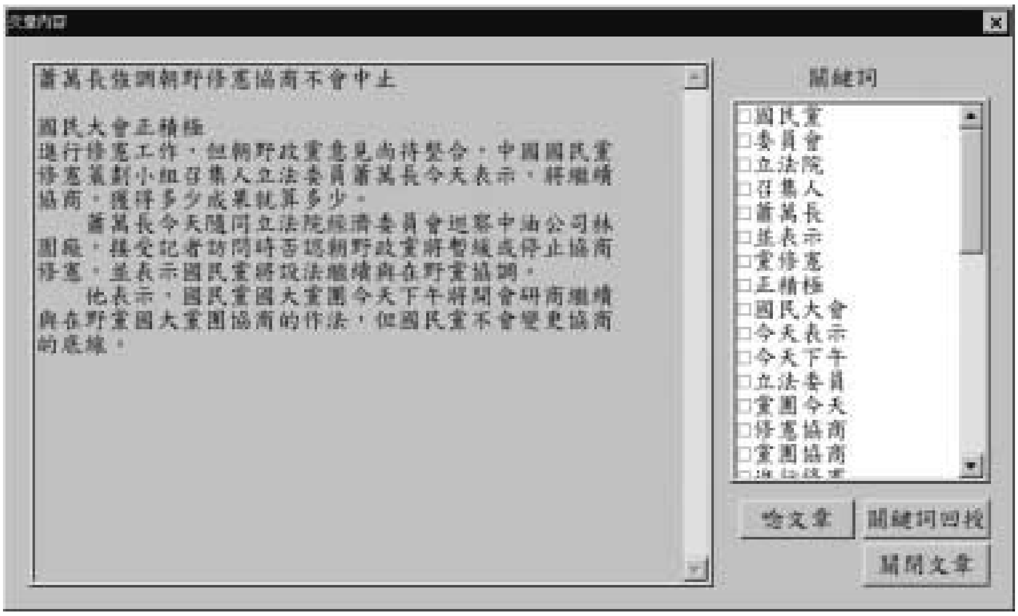
蕭萬長

```
( 0.:: 14)ㄍㄨㄛ_  ㄨㄛ_  ㄉㄨㄛ_  ㄏㄨㄛ_  ㄌㄨㄛ_  ㄖㄨㄛ_  ㄊㄨㄛ_  ㄋㄨㄛ_  ㄕㄨㄛ_  ㄓㄨㄛ_  ㄙㄨㄛ_
( 1.:: 26)ㄒㄧㄤ_  ㄐㄧㄤ_  ㄒㄧㄠ_  ㄑㄧㄤ_  ㄒㄧㄡ_  ㄒㄧㄚ_  ㄒㄧㄢ_  ㄕㄤ_  ㄒㄧㄥ_  ㄐㄧㄠ_  ㄑㄧ
( 2.:: 34) ㄔㄚ_  ㄘㄚ_  ㄗㄚ_  ㄙㄚ_  ㄔㄢ_  ㄕㄚ_  ㄔㄞ_  ㄓㄚ_  ㄔㄤ_  ㄊㄚ_  ㄘㄢ
( 3.:: 26)ㄒㄧㄠ_  ㄒㄧㄚ_  ㄒㄧㄡ_  ㄑㄧㄠ_  ㄐㄧㄠ_  ㄕㄠ_  ㄑㄧㄚ_  ㄐㄧㄚ_  ㄕㄚ_  ㄒㄧㄤ_  ㄓㄠ
( 4.:: 23) ㄨㄢ_  ㄍㄨㄢ_  ㄨㄣ_  ㄏㄨㄢ_  ㄉㄨㄢ_  ㄨㄞ_  ㄍㄨㄣ_  ㄖㄨㄢ_  ㄎㄨㄢ_  ㄉㄨㄢ_  ㄕㄨㄢ
```

```
□  23 8.35e-001 蕭萬長就任記者會.
□  20 7.33e-001 蕭萬長發表談話。
□  21 5.79e-001 馬國總理過境台北蕭萬長赴機場會晤
□  82 5.50e-001 行政院院會。
□ 468 4.96e-001 國民黨李主席約見黨內大老。
□ 291 4.60e-001 英國空軍結束旅駐香港任務
□  12 4.53e-001 國民黨中委選舉宋楚瑜最高票。
□  19 4.51e-001 蕭萬長宣示打擊犯罪決心。
□  22 4.16e-001 行政院長蕭萬長談柯江高峰會。
□ 265 3.90e-001 金門夏季缺水找內大陸探詢通水可行
□ 140 3.42e-001 台中科博館元旦特展
□  15 3.30e-001 連戰結束跨洲之旅返國。
□ 218 3.24e-001 屏東縣頂雙化廠動工遭抗爭
□ 126 3.16e-001 新竹少監典獄長遭彈劾
□  83 3.04e-001 行政院記者會談人質案。
□ 101 3.03e-001 立法院今天開始修訂兩岸合一法案
```

```
□1075 1.70e+000 蕭萬長否認國民黨與凍省只考輔簡單
□ 144 1.57e+000 李登輝約見蕭萬長提示黨對黨談判原
□1071 1.32e+000 蕭萬長強調修憲不會因出現僵局停止
□ 448 1.26e+000 蕭萬長重申爭取基層選票決心但時機
□ 750 1.25e+000 關於易人傳聞蕭萬長表無奈連戰指未
□ 751 1.24e+000 蕭萬長說關於同意權仍要與在野黨談
□ 883 1.22e+000 蕭萬長說黨對黨談判必須視黨團協商
□ 612 1.15e+000 蕭萬長強調朝野修憲協商不會中止
□1167 1.09e+000 蕭萬長為國民黨將繼續與在野黨協商
□ 479 9.31e-001 國新兩黨修憲談判雙方各派九代表
□ 190 7.37e-001 兩院制修憲提案加溫
□ 807 6.56e-001 因應九七陸委會將加速完成相關法制
□ 557 6.05e-001 因應七一政府相關活動初步定案
□ 260 5.80e-001 報告建議防範中共透過香港風俗敏感
□ 690 5.57e-001 陸委會諮詢委員為中共會透過香港對
□ 197 5.50e-001 修憲法條配套設計黨對黨談判仍漫長
```

Figure 3: The prototype Chinese text and speech information retrieval system.



蕭萬長強調朝野修憲協商不會中止

國民大會正積極
進行修憲工作，但朝野政黨意見尚待整合，中國國民黨
修憲策劃小組召集人立法委員蕭萬長今天表示，將繼續
協商，獲得多少成果就算多少。
　　蕭萬長今天隨同立法院經濟委員會巡察中油公司林
園廠，接受記者訪問時否認朝野政黨將暫緩或停止協商
修憲，並表示國民黨將設法繼續與在野黨協調。
　　他表示，國民黨國大黨團今天下午將開會研商繼續
與在野黨國大黨團協商的作法，但國民黨不會變更協商
的底線。

```
□國民黨
□委員會
□立法院
□召集人
□蕭萬長
□並表示
□黨修憲
□正積極
□國民大會
□今天表示
□今天下午
□立法委員
□黨團今天
□修憲協商
□黨團協商
```

Figure 4: The window to show a retrieved text record and the keywords extracted from the record.

respectively. Here a record title is used to represent each record, in form of either text or speech. There is a check box before each record title so that users can select relevant records by mouse and push the "相關回授 (relevance feedback)" button for further retrieval using the selected records.

On the other hand, to read or listen to the content of a retrieved record, the user only has to double-click the

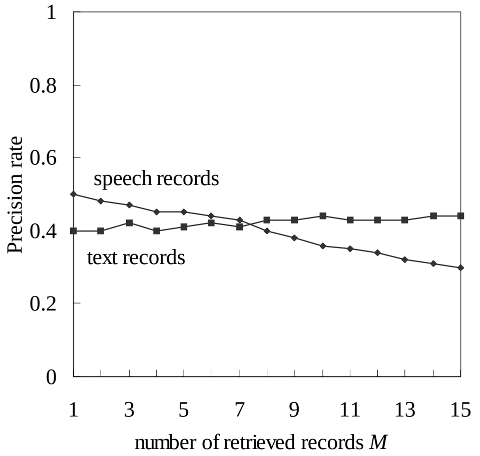


Figure 5: The precision rate with respect to the number of retrieved records *M*

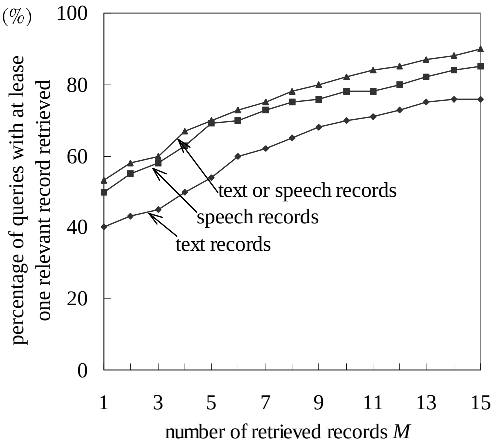


Figure 6: The percentage of queries with at lease one relevant record retrieved within top *M* records

record by mouse. If a voice record is double-clicked, the system will simply play the content of the voice record. If a text record is double-clicked, a popup window as shown in Figure 4 will be immediately created listing the content of the text record, with some keywords automatically extracted from the record also shown in the right sub-window. The user can select any relevant keywords and push the "關鍵詞回授(feedback by keywords)" button for relevance feedback using keyword information. Besides, the "唸文章(read the article)" button provides the function of reading the content of the text using a speech synthesis module.

4.1 System Performance

The system has been tested by 15 speakers with a total of 326 queries. Among these 326 queries, about 48 and 32 queries do not have any relevant records in the speech and text databases, respectively. Thus 278 and 294 queries are used as the measurement sets for speech database retrieval and text database retrieval. Since it is impossible to manually select all records relevant to each on-line tested query, the recall rates cannot be obtained here. Thus instead of recall-precision graph, here the precision rates with respect to the numbers of retrieved records are used as the measure for the system performance. The results for retrieving the voice and text records are shown in Figure 5. Since the result sub-windows in Figure 3 show only up to 15 voice records and 15 text records each time, here *M* is set from 1 up to 15. It can be found in Figure 5 that the curve for retrieving text records is much more flat than that for retrieving voice records. Because of the relatively rough information available for the voice records as compared to the text records, more candidate records must be selected to include some extra relevant records. Therefore more precision rate reduction is the necessary price paid to achieve higher recall rate while retrieving the voice records, which results in high slopes in the curves. Moreover, because the text database is much larger than the speech database, it is possible to retrieve more relevant text records at larger *M* with similar precision rates. However, very often there does not exist many voice records relevant to a given query, it is simply impossible to retrieve more relevant records even more candidate records are considered. In general, text information retrieval performs better than speech information retrieval. However, it can be found that, for smaller *M*, the retrieval of voice records even performs better than the retrieval of text records. One possible reason is that the rough information of syllable lattices on both sides for speech information retrieval may be more error-tolerant as compared to text information retrieval, in which the precise information is available on the database side.

Another performance measure used here is the percentage of queries for which at least one relevant record is retrieved within the top *M* selected records. For example, if 8 queries are tested and the first relevant records retrieved by these queries are ranked 2, 3, 6, 1, 10, 2, 8, 11, than we can obtain 12.5%, 37.5%, 50.0%, 50.0%, 50.0%, 62.5%, 62.5%, 75.0%, 75.0% and 87.5% of queries with at least one relevant records retrieved when top 1 to top 10 records are considered respectively. This measure shows not only the capability that the system can support at least one relevant record for each database within top *M* records, it also represents the capability that the system can support relevant records for users to select for relevance feedback. The results for retrieving the voice records and those for retrieving the text records are shown in Figure 6. It can be found that when 15 retrieved records are considered, 85% of queries will retrieve at least one relevant voice record and 76% of queries will retrieve at least one relevant text record. Since both retrieved voice records and text records can be selected and fedback to further retrieve both databases, Figure 6 also shows the percentage of queries with at least one relevant text or voice record retrieved within top *M* records for each database. It can be found that over 90% of queries can retrieve at least one relevant text or voice records within 15 records for each database.

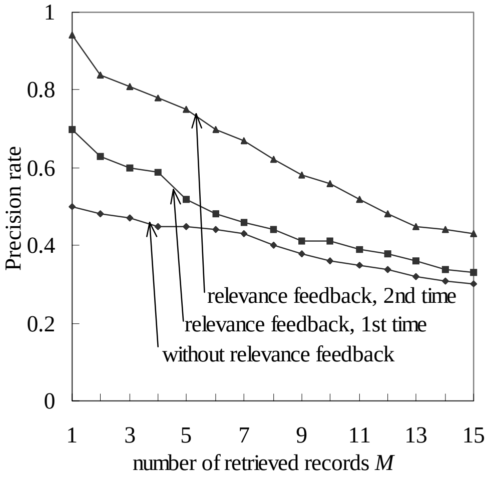


Figure 7: The performance of the relevance feedback approach for retrieving speech records measured by precision rate with respect to the number of retrieved records *M*

The performance of relevance feedback was tested with 189 queries out of the total of 326 testing queries mentioned above. The relevance feedback function in the system adopts both the relevance measure adjustment and query expansion schemes [8]. The precision rates with respect to the top *M* retrieved records for the tests with and without relevance feedback are shown in Figures 7 and 8 for speech and text database retrieval, respectively. Since the records fedback can always be retrieved and ranked very high, the measure of percentages of queries with at lease one relevant record retrieved within top *M* records is not shown again since it is almost 100% for the top 1 records. It can be found from Figures 7 and 8 that with the relevance feedback applied, the results for retrieving text records are much better than that for retrieving voice records and the curves for retrieving text records keep flat for larger number of retrieved records. This is also because the text database is much larger than the speech database. It is possible to retrieve more relevant text records, with similar precision rates at larger *M*, by the relevance feedback. On the other hand, very often there does not exist many voice records relevant to a given query, it is simply impossible to retrieve more relevant records even more candidate records are considered.

5. CONCLUSION

This paper deals with the problem of Chinese text and Mandarin speech information retrieval with Mandarin speech queries. Based on text keyword extraction, we have developed a new approach that can properly integrate the advantages of the syllable and word (keyword) information for retrieval. In addition, with the relevance feedback techniques, a prototype system with an interface supporting some user-friendly functions was successfully implemented. This prototype system not only verified the feasibility of the technologies that have been developed, but also provides a very good environment for further investigation on a variety of research topics on information retrieval.

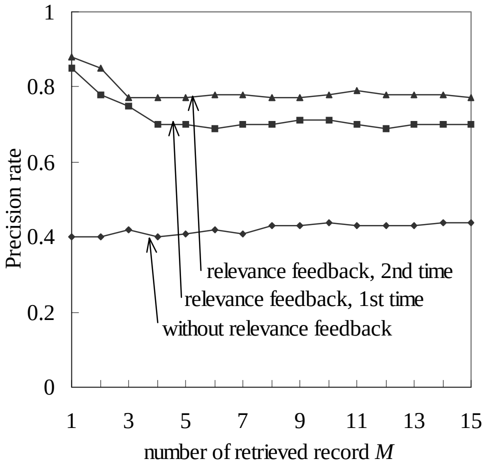


Figure 8: The performance of the relevance feedback approach for retrieving text records measured by precision rate with respect to the number of retrieved records *M*

REFERENCES

[1] K. Sparck Johns, G. J. F. Johns, J. T. Foote, and S. J. Young, “Experiments on Spoken Document Retrieval”, Information Processing & Management, Vol. 32, No. 4, pp. 399-417, 1996.

[2] Kenney Ng and Victor W. Zue, “Phonetic Recognition for Spoken Document Retrieval”, ICASSP, pp. 325-328, 1998.

[3] Bo-ren Bai, Chun-liang Chen, Lee-feng Chien, and Lin-shan Lee, “Intelligent Retrieval of Dynamic Networked Information from Mobile Terminal Using Spoken Natural Language Queries”, IEEE Transactions on Consumer Electronics, Vol. 44, No. 1, pp. 62-72, February 1998.

[4] Berlin Chen, Hsin-min Wang, Lee-feng Chien, and Lin-shan Lee, “A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification”, ICSLP, 1998.

[5] Sung-chien Lin, Lee-feng Chien, Keh-jiann Chen and Lin-shan Lee, “Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains”, Eurospeech, pp. 1203-1206, 1995.

[6] Bo-ren Bai, Lee-feng Chien, and Lin-shan Lee, “Very-Large-Vocabulary Mandarin Voice Message File Retrieval Using Speech Queries”, ICSLP, pp. 1950-1953, 1996.

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

[8] Lin-shan Lee, Bo-ren Bai, and Lee-feng Chien, “Syllable-based Relevance Feedback Techniques for Mandarin Voice Record Retrieval Using Speech Queries”, ICASSP, pp. 1459-1462, 1997.