SYLLABLE-BASED CHINESE TEXT/SPOKEN DOCUMENT RETRIEVAL USING TEXT/SPEECH QUERIES

BO-REN BAI
Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.

BERLIN CHEN* and HSIN-MIN WANG†
**Department of Computer Science & Information Eng.*
National Taiwan University
Institute of Information Science, Academia Sinica
Taipei, Taiwan, R.O.C.

In light of the rapid growth of Chinese information resources on the Internet, this study investigates a novel approach that deals with the problem of Chinese text and spoken document retrieval using both text and speech queries. By properly utilizing the monosyllabic structure of the Chinese language, the proposed approach estimates the statistical similarity between the text/speech queries and the text/spoken documents at the phonetic level using the syllable-based statistical information. The investigation successfully implemented a prototype system with an interface supporting some user-friendly functions and the initial test results demonstrate the feasibility of the proposed approach.

Keywords: Information retrieval; text document retrieval; spoken document retrieval; speech query; multi-modality; syllable lattice; speech recognition; Mandarin Chinese.

1. INTRODUCTION

Network technology and the Internet are creating a completely new information era. As widely anticipated, numerous digital libraries and multimedia databases will be available via the Internet in the near future. The digital libraries and multimedia databases will consist of heterogeneous types of information including text, audio, image and video. Intelligent and efficient information retrieval techniques allowing easy access to the extensive and varied information become highly desirable. With advances in speech recognition technology, many researchers have considered proper integration of information retrieval and speech recognition.[1,4,6,8,10,13] Since the Chinese language is nonalphabetic and the input of Chinese characters into computers is rather difficult, a multimodal interface for retrieving Chinese text/spoken documents is highly desired. Thus, this study focuses mainly on developing this framework.

Text retrieval has been studied for decades, while research on spoken document retrieval has only just begun. Unlike text documents, spoken documents cannot be retrieved by directly comparing them with speech queries. Both the speech

†Author for correspondence. E-mail: whm@iis.sinica.edu.tw

International Journal of Pattern Recognition and Artificial Intelligence, Vol. 14, No. 5 (2000) 603–616

queries and spoken documents must be converted into content features such as keywords, phone strings, and texts using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can be measured. Selecting appropriate content features to represent the spoken documents and speech queries is thus very important. Meanwhile, these features must also be appropriate for text document retrieval, so that the same strategy can be applied to retrieval of both text and spoken documents. Given the characteristic monosyllabic structure of the Chinese language, this study investigates a syllable-based approach, which conducts the statistical similarity estimation between the text/speech queries and the text/spoken documents at the phonetic level using the syllable-based statistical information. Although more than 10 000 commonly used Chinese characters exist, each character is monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1345. The combination of these monosyllabic characters, or 1345 syllables, gives an almost unlimited number of monosyllabic or polysyllabic Chinese words. Mandarin Chinese is a tonal language, in which each syllable is assigned a tone, chosen from a total of four lexical tones plus one neutral tone. If the differences among the syllables caused by tones are ignored, then only 416 base syllables (i.e. the syllable structures independent of the tones) instead of 1345 tonal syllables are required to cover the pronunciations for Mandarin Chinese. The monosyllabic nature of the Chinese language makes it feasible to measure the similarity between the text/spoken documents and the text/speech queries directly at the syllable level. This approach bypasses the high ambiguity caused by the one-to-many mapping relation from syllables to characters and significantly reduces the computational requirements. Furthermore, in the Chinese language, many loan words derived from foreign languages are proper names or technical terms, and these words are important terms for information retrieval. However, a foreign word can very often be translated into different Chinese words, for example "Kosovo" may be translated into "科索佛", "科索夫", and "科索伏", and "柯索佛" this diversity will cause serious retrieval errors in a character-based or word-based Chinese information retrieval system. The syllable-based approach can overcome this problem since all varieties of a given loan word usually share a syllable string approximating the pronunciation of the original foreign word, despite consisting of different character strings. Because text/spoken documents and text/speech queries share the same syllable-based features, the consistency of the whole system can be preserved.

The rest of this paper is organized as follows. Section 2 introduces the overall architecture of the proposed approach for Chinese text/spoken document retrieval. After briefly reviewing in Sec. 3, the continuous Mandarin speech recognition technology, Secs. 4 and 5 then describe the feature vector and the retrieving process used in the proposed approach. Then, Sec. 6 presents some experimental results and Sec. 7 outlines the prototype system. Concluding remarks are finally made in Sec. 8.

2. OVERALL ARCHITECTURE OF CHINESE TEXT/SPOKEN DOCUMENT RETRIEVAL

Figure 1 displays the overall architecture of the proposed approach for Chinese text/spoken document retrieval. The whole system can be divided into three parts. The first part, located in the upper dotted square in Fig. 1, is the offline processing subsystem. All the processes in this part should be performed offline in advance. The second part, located in the middle dotted square, is the initialization subsystem. All the processes in this part should be performed in the system initialization stage. The third part, located in the lower dotted square, is the online retrieval subsystem, in which all the processes must be performed online in real-time. The detailed operations of each part are described individually below.

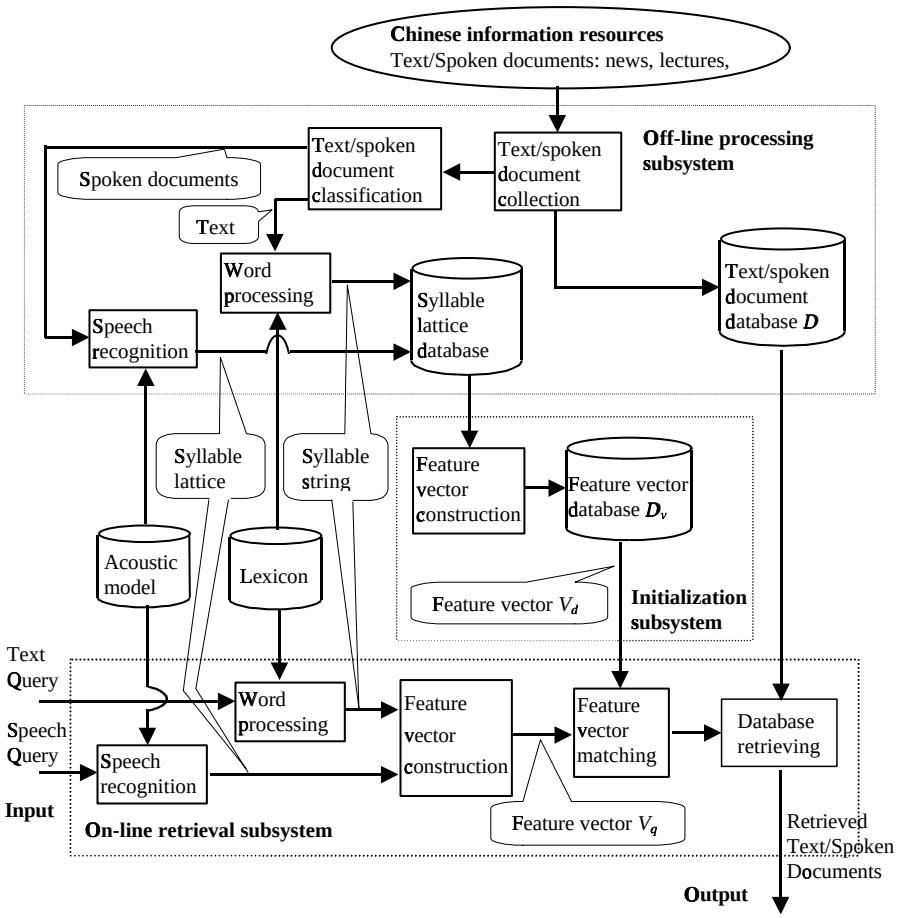


Fig. 1. Overall architecture of the proposed approach for Chinese text/spoken document retrieval. The input can be both text and speech queries, and these queries can be quasi-natural-language queries or simple keyword queries, while the output includes both text and spoken documents.

In the offline processing subsystem, the collected documents can be both text documents and spoken documents. For a spoken document, speech recognition is first applied to generate a syllable lattice, including the acoustic scores for all syllable candidates, and the syllable lattice is then added to the syllable lattice database. On the other hand, for a text document, word processing is applied to perform word segmentation[3] and phonetic labeling according to a general Chinese lexicon[5] to produce a syllable string. Herein, the syllable string can be considered as a syllable lattice with only the top one candidate and, thus, it can also be stored in the syllable lattice database. The lexicon used in this study contains 14 052 single character words and another 70 000 words composed from two to four characters. In this way, the most time-consuming speech recognition process is performed offline in advance, and all information necessary for retrieval is stored in the syllable lattice database.

The initialization subsystem extracts the feature vectors to be used for retrieval from the syllable lattice database. The feature vector of each document contains the presence information, frequency counts, and acoustic scores of all the syllables and adjacent syllable pairs in the syllable lattice. After the feature vectors have been created for all syllable lattices in the syllable lattice database, the feature vector database D_v is established, which will be the target database for physical retrieval. The whole process is also performed offline in advance.

In the online retrieval subsystem, when a speech query is entered, speech recognition first generates a syllable lattice for the speech query and, then, the corresponding feature vector V_q will be constructed based on this syllable lattice via an identical processing procedure as that for spoken documents. For a text query, word processing first generates the corresponding syllable string and, then, the feature vector V_q will be constructed based on this syllable string via an identical processing procedure as that for text documents. Given feature vector database D_v and query feature vector V_q, the retrieving module then evaluates the similarity between V_q and all the feature vectors in the database, and selects the most similar set of documents as the retrieval output. Section 5 discusses the retrieval process in further detail.

3. SYLLABLE RECOGNITION OF CONTINUOUS MANDARIN SPEECH

As mentioned earlier, Mandarin Chinese contains 1345 phonologically allowed tonal syllables, reducible to 416 base syllables and five tones. Base syllable recognition is, thus, considered the first key problem in spoken document retrieval for Mandarin Chinese. However, although the base syllable is a natural recognition unit for Mandarin Chinese due to the monosyllabic structure of the Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computational requirements in the recognition phase. Thus, the acoustic units chosen herein are context-dependent Initial/Finals,[14] chosen due to the monosyllabic nature of Mandarin Chinese and the Initial/Final structure of Mandarin base syllables.

Here, Initial refers to the initial consonant of the base syllable, and Final refers to the vowel (or diphthong) part but including optional medial or nasal ending. Each Initial or Final is then represented by a left-to-right continuous HMM.[11] To facilitate ease of use, the retrieval system is operated under the speaker-independent mode. That is, the speaker-independent context-dependent Initial/Final HMM's are used to recognize the syllables and construct the syllable lattices. These models are trained by using a training speech database including 5.3 hours of speech of phonetically balanced sentences and isolated words produced by roughly eighty male and forty female speakers. Also, to deal with the silence segments in the spoken documents or speech queries, the investigation uses a single state HMM to represent the silence.

Based on the acoustic models mentioned above, the speech recognition procedure for a spoken document is described as follows. In the first pass, the speech recognizer performs the Viterbi search on the spoken document and outputs the best syllable sequence and the corresponding syllable boundaries. In the second pass, based on the state likelihood scores calculated in the first pass search and the syllable boundaries of the best syllable sequence, the speech recognizer performs the Viterbi search on each utterance segment that may contain a syllable and outputs several most likely syllable candidates along with their acoustic recognition scores. Then, after the two-pass speech recognition process is complete, a syllable lattice can be constructed.

Because the original acoustic recognition score $\log p(O|s)$ for a certain syllable candidate s with a certain utterance segment O can be negative or positive, the score should be further manipulated so that it can be used in retrieval. In this study, $\log p(O|s)$ is transformed into a range between 0 and 1 by a Sigmoid function. That is, the final acoustic recognition score $\mathrm{as}(s)$ can be obtained using the following equation,

$$\mathrm{as}(s) = \frac{2}{1 + \exp(-\alpha \times (\log p(O|s) - \log p(O|s^*)))} \tag{1}$$

where $\log p(O|s^*)$ is the original acoustic recognition score of the top 1 syllable candidate, and α is used to control the slope of the Sigmoid function. In the following experiments, α is set to 0.1. Based on Eq. (1), the final acoustic recognition score of the top 1 syllable candidate is always fixed as 1 and the final acoustic recognition scores of the other syllable candidates are transformed to a range between 0 and 1.

An identical procedure can be applied to speech queries to generate the corresponding syllable lattices. Of course, the syllable lattice construction procedure is performed offline in advance for spoken documents, but online in real-time for speech queries.

4. FEATURE VECTORS

Before detailing the retrieving process, this section introduces the feature vectors used for the similarity measure. For each document d in the database D, through

searching the syllable lattice, all the acoustic scores of single syllables and adjacent syllable pairs in the syllable lattice can be extracted to form the feature vector V_d,

$$V_d = (s(s_1), \ldots, s(s_i), \ldots, s(s_{416}), s(s_1, s_1), \ldots, s(s_i, s_j), \ldots, s(s_{416}, s_{416})) \quad (2)$$

where $s(s_i)$ and $s(s_i, s_j)$ are the scores of syllable s_i and syllable pair (s_i, s_j), respectively. Meanwhile, $s(s_i)$ is the sum of the acoustic recognition scores, $\mathrm{as}(s_i)$, of all the occurrences of syllable s_i in the syllable lattice. The acoustic recognition score of syllable pair (s_i, s_j) is simply the sum of the acoustic recognition scores of its component syllables, that is, $\mathrm{as}(s_i, s_j) = \mathrm{as}(s_i) + \mathrm{as}(s_j)$, and $s(s_i, s_j)$ is the sum of the acoustic recognition scores, $\mathrm{as}(s_i, s_j)$ of all the occurrences of adjacent syllable pair (s_i, s_j) in the syllable lattice. For a text document, this investigation uses the frequency counts instead of the acoustic score information to form the feature vector. The feature vector construction procedure can be performed offline on all documents in the database D to form a feature vector database D_v, which will be the target database for physical retrieval.

Generally, the syllable information is fairly rough while the syllable pair information is more precise. However, the syllable information is very crucial to Chinese text/spoken document retrieval. In Mandarin Chinese, many proper names, often the key information for retrieval, are commonly abbreviated, such as "中央研究院 (Academia Sinica)" abbreviated as "中研院", "台灣大學 (National Taiwan University)" as "台大", and so on. For each example, the adjacent syllable pairs obviously do not match at all, despite representing the same organization. Thus, it is impossible to obtain documents containing "中央研究院" using the query "中研院", and vice versa. Furthermore, it is also obvious that the word order in Mandarin Chinese is frequently flexible, a good example is "李遠哲院長" and "李院長遠哲" which both represent "Lee Yuan-Tseh, the president of Academia Sinica". Both contain the identical five syllables, but only two of the four adjacent syllable pairs are the same, specifically, "院長" and "遠哲". Additionally, when speech recognition errors, such as deletions, insertions, and substitutions, occur in speech queries or spoken documents, the desired documents may be unretrievable if only the syllable pair information is used. Here is an example of the worst case. If the input query "中研院" is recognized as "中院" or "中 X 院" (where X represents a substitution error), both syllable pairs "中研" and "研院" in "中研院" will be lost, and the documents containing "中研院" will not be retrieved. On the other hand, if the same recognition errors occur in a spoken document which contains "中研院", then it will not be retrieved by the query "中研院", either. Therefore, this investigation incorporates both syllable information and syllable pair information in the feature vector.

With respect to a query, the same feature vector construction procedure must be performed online to construct the feature vector V_q immediately after the input query is entered. On the other hand, to reduce ambiguity caused by the irrelevant words that frequently appear in quasi-natural-language queries, such as "我想要找 ... (I would like to find ...)" and "有沒有關於 ... (Is there anything

about ...)", the inverse document frequency,[12] which has been widely adopted in many conventional text information retrieval systems, is applied to the feature vector V_q.

5. RETRIEVING PROCESS

Given the feature vector database D_v and a query q, the retrieval problem becomes a search process to retrieve the document d^* in target database D_v which is most related to the query. This search process can be formulated as follows:

$$d^* \equiv \arg\max \mathrm{Sim}(d, q) \tag{3}$$

where $\mathrm{Sim}(d, q)$ is a similarity measure between a document d and a query q. In this study, a Cosine measure[12] is used to estimate the similarity:

$$\mathrm{Sim}(d, q) = \cos(V_d, V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|}\,. \tag{4}$$

In this way, a larger $\mathrm{Sim}(d, q)$ value implies a more relevant document d to the query q. Documents with larger $\mathrm{Sim}(d, q)$ values will thus be selected and ranked as the retrieval results.

6. EXPERIMENTS AND DISCUSSIONS

This section presents several experiments to show the feasibility of the above approaches. The database used in the following experiments is described first.

6.1. Database Used in the Experiments

The example database used for simulation experiments consists of 500 Chinese text documents and 500 Mandarin spoken documents. The text materials are news articles published in the Taiwan area in 1997. The spoken documents were produced by five male speakers, and each speaker read 100 of the 500 text documents as a news announcer does. On average, each document contains about 100 characters (i.e. 100 syllables), and their length ranges from 44 to 269 characters. Meanwhile, 160 speech queries produced by four male speakers were used for testing, and they were further manually transcribed into text queries. Eighty of the queries are simple queries, each containing only one key phrase for some news item without any irrelevant words. An example phrase is "亞太經合會", which is a common abbreviation of "亞洲太平洋經濟合作會議(Asia Pacific Economic Cooperation, APEC)". The other 80 queries, which contain some irrelevant words in addition to the key phrases, are quasi-natural-language queries of the above 80 simple queries. For example, "有沒有關於亞太經合會的新聞? (Is there any news about APEC?)". For the 500 spoken documents (or 500 text documents), the documents relevant to each query were manually identified in advance for performance evaluation purposes. Each query has on average 5.9 relevant documents from the 500 documents in the database, with the exact number ranging from 1 to 20.

For each spoken document and speech query, speech recognition was applied to generate a corresponding syllable lattice. On the other hand, all the text documents and text queries were automatically labeled to obtain their corresponding syllable strings. The 500 syllable lattices and 500 syllable strings comprise the whole syllable lattice database used in the following experiments. Although the text and spoken documents of a practical text/spoken database generally contain different contents, the same materials used herein facilitate an evaluation of the degree of difficulty among the four different categories of retrieval problems.

6.2. Chinese Text/Spoken Document Retrieval Using Simple Queries

The first experiment aimed to show the performance of retrieving Chinese text/spoken documents using simple text/speech queries. Figure 2 presents the results in the recall-precision graph,[7] where TQ, TD, SQ and SD represent the text queries, text documents, speech queries and spoken documents, respectively. The curve marked by "TQ/TD" indicates the results of using text queries to retrieve text documents, and so on. Among the four categories of retrieval, using text queries to retrieve text documents achieves the best performance, while using speech queries to retrieve spoken documents is the most difficult task, due to the possibility of recognition errors occurring in both speech queries and spoken documents. Compared with the results of using text queries to retrieve text documents, a large performance degradation was found for the other three categories of retrieval that speech recognition must be applied in either queries or documents, or both. Using speech queries to retrieve spoken documents clearly works worst. This is reasonable since, in this case, both the information to be retrieved and the input queries are in the form of speech rather than text, with unknown variability on both sides. However, in the other two categories, relatively precise information is available on one side. Also, using speech queries to retrieve text documents produces slightly better performance than using text queries to retrieve spoken documents. In fact, the noninterpolated average precision[7] of these two categories of retrieval is 0.63 and 0.58, respectively. This result is because automatically recognizing the longer spoken documents is much more difficult than the relatively short input speech queries. Interestingly, the two curves for text document retrieval are much flatter than those for spoken document retrieval. For spoken document retrieval, only rough information is available on the document side because of the recognition errors and, thus, more documents are needed to include all relevant documents. Restated, the cost of the high recall rate is a significantly reduced precision.

6.3. Chinese Text/Spoken Document Retrieval Using Quasi-Natural-Language Queries

This experiment attempted to evaluate the performance of retrieval using quasi-natural-language queries. Figure 3 displays the results in the recall-precision graph, revealing very similar trends to those discussed in the previous experiment for simple queries. An exception is that the performance difference between using speech

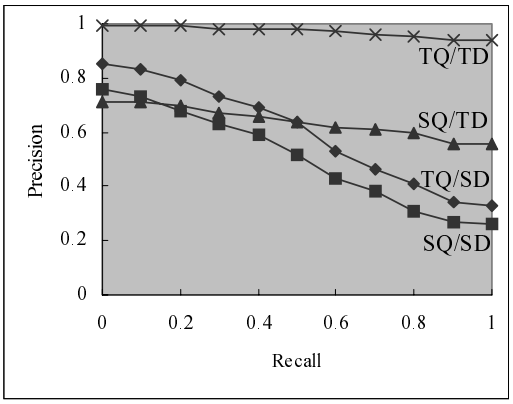


Fig. 2. Performance of Chinese text/spoken document retrieval using simple queries. TQ/SD and SQ/SD represent spoken document retrieval with text queries and speech queries, respectively, while TQ/TD and SQ/TD represent text document retrieval with text queries and speech queries, respectively.

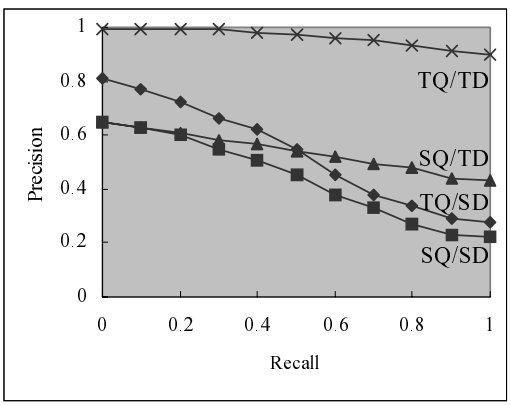


Fig. 3. Performance of Chinese text/spoken document retrieval using quasi-natural-language queries. TQ/SD and SQ/SD represent spoken document retrieval with text queries and speech queries, respectively, while TQ/TD and SQ/TD represent text document retrieval with text queries and speech queries, respectively.

queries to retrieve text documents and using text queries to retrieve spoken documents is less obvious here. In fact, their noninterpolated average precision is 0.53 and 0.52, respectively. This result is because automatic recognition of long quasi-natural-language queries is no longer as simple as automatic recognition of short simple queries. Furthermore, Fig. 4 illustrates the results for using simple queries and quasi-natural-language queries in four categories of retrieval together for comparison in noninterpolated average precision. For each category of retrieval, although simple queries naturally yield better results than quasi-natural-language queries, the difference is insignificant, especially when using text queries to retrieve text documents. The above results indicate that the inverse document frequency in-

formation used here can partially reduce the ambiguity caused by irrelevant words, such as "我想要找..." ("I would like to find ...") and "有沒有關於..." ("Is there anything about ..."), contained in quasi-natural-language queries. However, further improvements are needed.

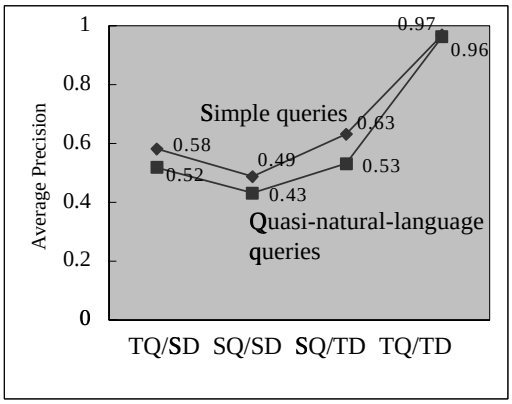


Fig. 4. Comparison between using simple queries and quasi-natural-language queries. TQ/SD and SQ/SD represent spoken document retrieval with text queries and speech queries, respectively, while TQ/TD and SQ/TD represent text document retrieval with text queries and speech queries, respectively.

7. THE PROTOTYPE SYSTEM

This investigation has successfully implemented a prototype system for Chinese text and speech information retrieval, as shown in Fig. 5. The system contains an interface supporting some user-friendly functions. The upper left subwindow lists the keyword vocabulary used in the system. These keywords were automatically extracted offline in advance from the text database. The details of keyword extraction can be found in Ref. 2. The subwindow right of the keyword list displays the speech waveform of the input query. Additionally, several buttons including "離開 (exit)", "調適 (adaptation")", "開始檢索" (begin to retrieve)" and "相關回授 (relevance feedback)" exist for corresponding functions. The details of relevance feedback techniques used in the prototype system can be found in Ref. 9. The upper right subwindow displays the keywords spotted from the input speech query by the speech keyword spotter. Meanwhile, the middle subwindow below the above upper subwindows shows the syllable lattice constructed by the continuous speech recognizer. Both the recognized keywords and the syllable lattice are used for retrieving text and speech databases, that is, the keyword scores and the similarity measure obtained by Eq. (4) are summed together as the final similarity measure between a document and a query. The lower left and lower right subwindows show the retrieved results of the speech database (語音資料庫) and text database (文字資料庫), respectively. The speech database consists of 500 documents while the text database consists of 5 819 documents. The contents of both are news published in Taiwan during different time spans. Herein, a document title is used to

represent each document, in the form of either text or speech. A check box exists before each document title so that users can select relevant documents by mouse and push the "相關回授 (relevance feedback)" button for further retrieval using the selected documents.

To read or listen to the content of a retrieved document, the user merely has to double-click on the document. If a spoken document is double-clicked, the system will play the content, while if a text document is double-clicked, a popup window will list the contents of the text document.

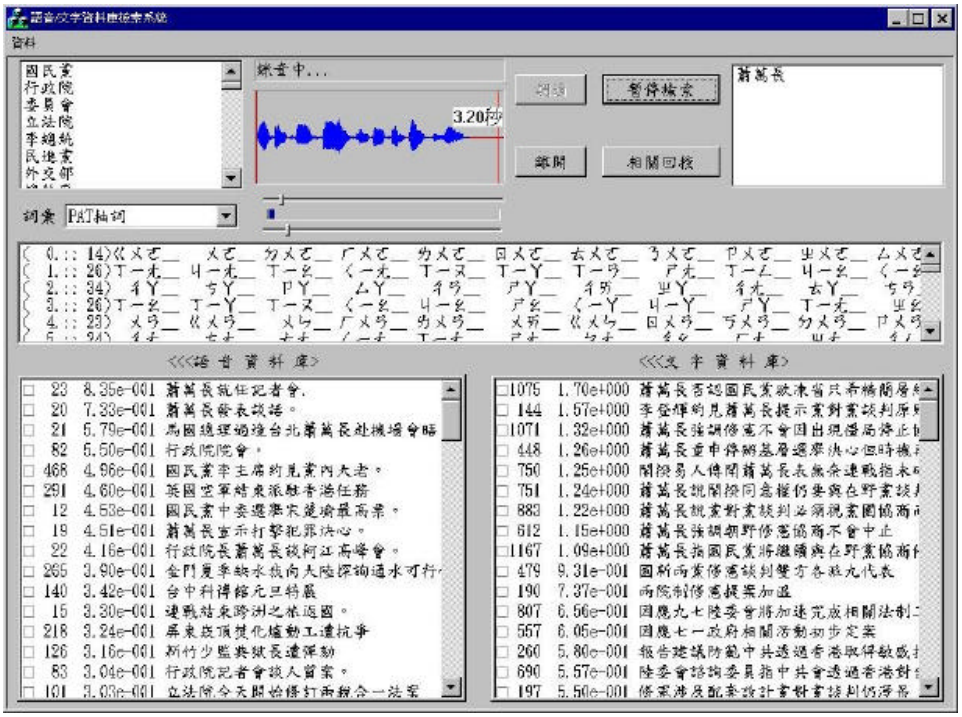


Fig. 5. The prototype Chinese text and speech information retrieval system.

7.1. System Performance

The system was tested by 15 speakers with 326 queries. Among these 326 queries, about 48 and 32 queries had no relevant documents in the speech and text databases, respectively. Thus 278 and 294 queries are used as the measurement sets for speech and text database retrieval, respectively. Since it is impossible to manually select all documents relevant to each online tested query, the recall rates cannot be obtained herein. Therefore, the performance measure used here is the percentage of queries for which at least one relevant document is retrieved within the top M selected documents. For example, if eight queries are tested and the first relevant documents retrieved by these queries are ranked 2, 3, 6, 1, 10, 2, 8, 11, then 12.5%, 37.5%, 50.0%, 50.0%, 50.0%, 62.5%, 62.5%, 75.0%, 75.0% and 87.5% of

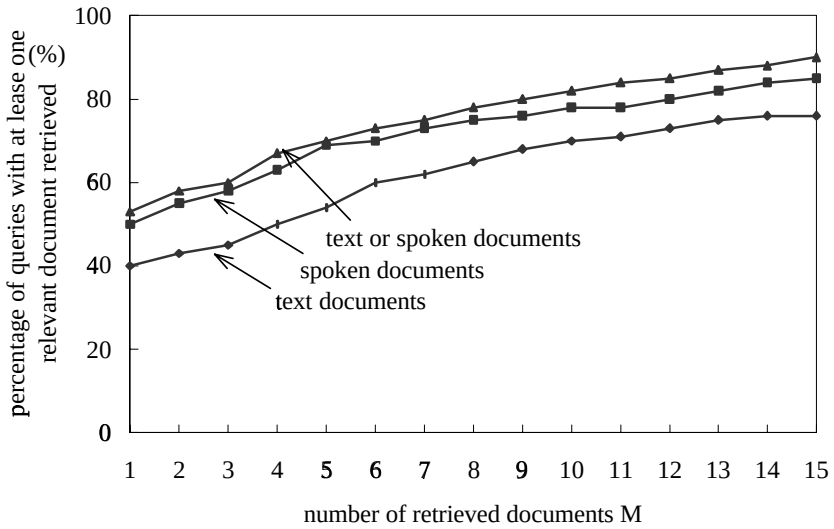


Fig. 6. Percentage of queries with at least one relevant document retrieved within the top M documents.

queries retrieve at least one relevant document when the top 1 to top 10 documents are considered respectively. This measure shows not only the capability that the system can support at least one relevant document for each database within the top M documents, it also represents the capability that the system can support relevant documents for users to select for relevant feedback. Figure 6 summarizes the results for retrieving the spoken documents and text documents. When 15 retrieved documents were considered, 85% of queries retrieved at least one relevant spoken document and 76% of queries retrieved at least one relevant text document. Since both retrieved spoken documents and text documents can be selected and fedback to further retrieve both databases, Fig. 6 also shows the percentage of queries with at least one relevant text or spoken document retrieved within the top M documents for each database. Over 90% of the queries retrieved at least one relevant text or spoken document within the first fifteen documents for each database.

8. CONCLUDING REMARKS AND FUTURE WORK

This paper describes the initial results of a long-term research project on speech retrieval. The popularity of the Internet and multimedia has made this area very important. This paper presents a syllable-based approach for retrieving Chinese text/spoken documents using both text and speech queries. Based on this approach, this investigation has successfully implemented a prototype system. The experimental results reported here are not necessarily satisfactory, and many problems undoubtedly remain unsolved. However, these preliminary results at least indicate the good potential of this direction.

Currently, the authors are trying to evaluate the proposed approach using a large real-world database of news broadcasts. Automatic recognition of such spoken

materials is a challenging task. Furthermore, the relevant feedback techniques, such as query expansion and term suggestion schemes, are currently being studied to enhance retrieval performance and human–computer interaction.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Lin-Shan Lee and Dr. Lee-Feng Chien for their valuable assistance and comments. The authors would also like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 87-2213-E-001-026 and NSC 88-2213-E-001-019.

REFERENCES

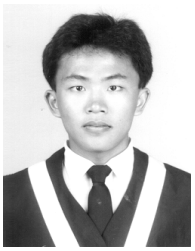
1. B. R. Bai, L. F. Chien and L. S. Lee, "Very-large-vocabulary Mandarin voice message file retrieval using speech queries," *ICSLP96*, pp. 1950–1953.
2. B. R. Bai, C. L. Chen, L. F. Chien and L. S. Lee, "Intelligent retrieval of dynamic networked information from mobile terminal using spoken natural language queries," *IEEE Trans. Consumer Electron.* **44**, 1 (1998) 62–72.
3. K. J. Chen and S. H. Liu, "Word identification for Mandarin Chinese sentences," *COLING92*, pp. 101–107.
4. L. F. Chien *et al.*, "Internet Chinese information retrieval using unconstrained Mandarin speech queries based on a client-server architecture and a PAT-tree-based language model," *ICASSP97*, pp. 1155–1158.
5. CKIP group, "Analysis of syntactic categories for Chinese," CKIP Technical Report, No. 93-05, Institute of Information Science, Academia Sinica, Taipei, 1993.
6. U. Glavitsch and P. Schäuble, "A system for retrieving speech documents," *ACM SIGIR Conf. R&D in Information Retrieval*, 1992, pp. 168–176.
7. D. Harman, *Overview of the Fourth Text Retrieval Conference* (*TREC-4*), available at "http://trec.nist.gov/pubs/trec4/t4_proceedings.html".
8. J. Kupiec, D. Kimber and V. Balasubramanian, "Speech-based retrieval using semantic co-occurrence filtering," *The Human Knowledge Technology Workshop*, 1994, pp. 373–377.
9. L. S. Lee, B. R. Bai and L. F. Chien, "Syllable-based relevance feedback techniques for Mandarin voice record retrieval using speech queries," *ICASSP97*, pp. 1459–1462.
10. K. Ng and V. Zue, "Subword unit representations for spoken document retrieval," *EUROSPEECH97*, pp. 1607–1610.
11. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International, Inc., 1993.
12. G. Salton and M. McGill, *Introduction to Modern Information Retrieval*, NY, McGraw-Hill, 1983.
13. K. Spärck Johns, G. J. F. Johns, J. T. Foote and S. J. Young, "Experiments on spoken document retrieval," *Inf. Process. Manag.* **32**, 4 (1996) 399–417.
14. H. M. Wang *et al.*, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data," *IEEE Trans. Speech Audio Process.* **5**, 2 (1997) 195–200.

Bo-Ren Bai received his B.S. degree and Ph.D. in electrical engineering from National Taiwan University in 1992 and 1998, respectively.

His research interests include speech processing and information retrieval.

Hsin-Min Wang received his B.S. degree and Ph.D. in electrical engineering from National Taiwan University in 1989 and 1995, respectively. Since 1996, he has been an assistant research fellow with the Institute of Information Science, Academia Sinica, Taipei.

His research interests include speech recognition, natural language processing, speech information retrieval and spoken dialogue.



Berlin Chen received his B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University in 1994 and 1996, respectively. Since 1996, he has been working with the Chinese Knowledge Information Processing Group at the Institute of Information Science, Academia Sinica, Taipei. He is currently pursuing a Ph.D. in computer science and information engineering at National Taiwan University.

His research has been focused on speech processing and natural language processing.