Syllable-Based Chinese Text/Spoken Document Retrieval Using Text/Speech Queries

Bo-ren Bai
Department of Electrical Engineering,
National Taiwan University
Taipei, Taiwan, Republic of China
e-mail: white@speech.ee.ntu.edu.tw

Berlin Chen and Hsin-min Wang
Institute of Information Science,
Academia Sinica
Taipei, Taiwan, Republic of China
e-mail: {berlin,whm}@iis.sinica.edu.tw

Abstract

In order to solve the problem with the fast growth of Chinese information resources on the Internet, this paper deals with the problem of Chinese text and spoken document retrieval using both text and speech queries. By properly utilizing the monosyllabic structure of Chinese language, the proposed approach performs the statistical similarity estimation between the text/speech queries and the text/spoken documents directly at the phonetic level using syllable-based statistical information. Extensive experiments were performed and the preliminary results are very encouraging.

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. With the advances in speech recognition technology, proper integration of information retrieval and speech recognition has been considered by many researchers [1-6]. Because Chinese language is not alphabetic and input of Chinese characters into computers is very difficult, a multi-modal interface for retrieving Chinese text/spoken documents is especially highly desired, and thus this framework is the primary focus of this paper.

Text retrieval has been investigated for decades while the research on spoken document retrieval has just begun. Unlike the text documents, the spoken documents can't be retrieved by directly comparing them with the speech queries. Both the speech queries and the spoken documents must be transcribed into some kind of content features such as keywords, phone strings, texts, and so on using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can be measured. To choose appropriate content features to represent the spoken documents as well as the speech queries is therefore very important. On the other hand, these features should be also appropriate for text document retrieval, such that exactly the same strategy can be applied to retrieval of both text and spoken documents. Considering the characteristic monosyllabic structure of Chinese language, this paper presents a syllable-based approach, which performs the statistical similarity estimation between the text/speech queries and the text/spoken documents directly at the phonetic level using syllable-based statistical information. Although there are more than 10,000 commonly used Chinese characters, each character is monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1,345. The combination of these monosyllabic characters, or 1,345 syllables gives almost unlimited number of monosyllabic or polysyllabic Chinese words. Also, 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, then only 416 base syllables (i.e., the syllable structures independent of the tones) instead of 1,345 different tonal syllables are required to cover the pronunciations for Mandarin Chinese. This special monosyllabic feature of Chinese language makes it feasible to measure the similarity between the text/spoken documents and the text/speech queries directly at the syllable level. In this way, the high ambiguity caused by the one-to-many mapping relation from syllables to characters in Mandarin speech recognition can be completely bypassed and the computation requirement can be significantly reduced. Furthermore, in Chinese language, many loan words that are derived from foreign languages are proper names or technical terms, and these words are important terms for information retrieval. However, a foreign word is very often translated into different Chinese words, e.g. "Clinton" may be translated into "柯林頓", or "克林頓",

and so on, and such diversity definitely will cause serious retrieving errors in a character-based or word-based Chinese information retrieval system. The syllable-based approach here can somehow bypass this problem since those loan words usually share the same syllable string that is similar to the pronunciation of the original foreign word, though they may consist of different character strings. Because the same syllable-based features are used for text/spoken documents and text/speech queries, the consistence of the whole system thus can be preserved.

The rest of this paper is organized as follows. The overall architecture of the proposed approach for Chinese text/spoken document retrieval is introduced in section 2. After briefly reviewing the continuous Mandarin speech recognition technology in section 3, the feature vector and the retrieving process used in our approach are then introduced in sections 4 and 5 respectively. Section 6 presents the experimental results. Finally, the concluding remarks are made in section 7.

2. Overall Architecture of Chinese Text/Spoken Document Retrieval

The overall architecture of the proposed approach for Chinese text/spoken document retrieval is shown in Figure 1. The whole system can be divided into three parts. The first part in the upper dotted square of Figure 1 is the off-line processing subsystem. All processes in this part should be performed off-line in advance. The second part in the middle dotted square is the initialization subsystem. All processes should be performed in the system initialization stage. The third part in the lower dotted square is the on-line retrieval subsystem, in which all processes must be performed on-line in real-time. The detailed operation of each part will be described separately below.

In the off-line 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. While for a text document, the word processing is applied to perform word segmentation [7] and phonetic labeling according to a general Chinese lexicon [8] to generate a syllable string. Here, the syllable string can be thought as a syllable lattice with only top 1 candidate and thus can be also stored in the syllable lattice database. The lexicon used in this paper contains 140,52 single character words and another 70,000 words composed of from 2 to 4 characters. In this way, the most time consuming speech recognition process is performed off-line in advance, and all information necessary for retrieval are stored in the syllable lattice database.

The initialization subsystem is to obtain 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 syllables and adjacent syllable pairs in the syllable lattice. After the feature vectors have been constructed for all syllable lattices in the syllable lattice database, the feature vector database D_v is established, which will be the target database to be physically retrieved. The whole process is also performed off-line in advance.

In the on-line retrieval subsystem, when a speech query is entered, speech recognition will 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 exactly the same processing procedures as those for spoken documents. For a text query, the word processing will generate the corresponding syllable string, and then the feature vector V_q will be constructed based on this syllable string via exactly the same processing procedures as those for text documents. Given the feature vector database D_v and the query feature vector V_q, the retrieving module then evaluates the similarity measure between V_q and all feature vectors of the database, and selects a set of documents with the highest similarity measures as the retrieving output. The further details of the retrieving process will be discussed below in section 5.

3. Syllable Recognition of Continuous Mandarin Speech

As mentioned above, in Mandarin Chinese, there exists a total of 1,345 phonologically allowed tonal syllables, and these tonal syllables can be reduced to 416 base syllables and 5 tones. Base syllable recognition is thus believed to be the first key problem for spoken document retrieval considered here. However, although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computation requirement in the recognition phase. Thus, the acoustic units chosen here are context-dependent Initial/Final's [9] specially considering the monosyllabic nature in Mandarin Chinese and the Initial/Final structure in 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 or Final is then represented by a left-to-right continuous HMM's. To allow anyone to use the system naturally without training, 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 a training speech

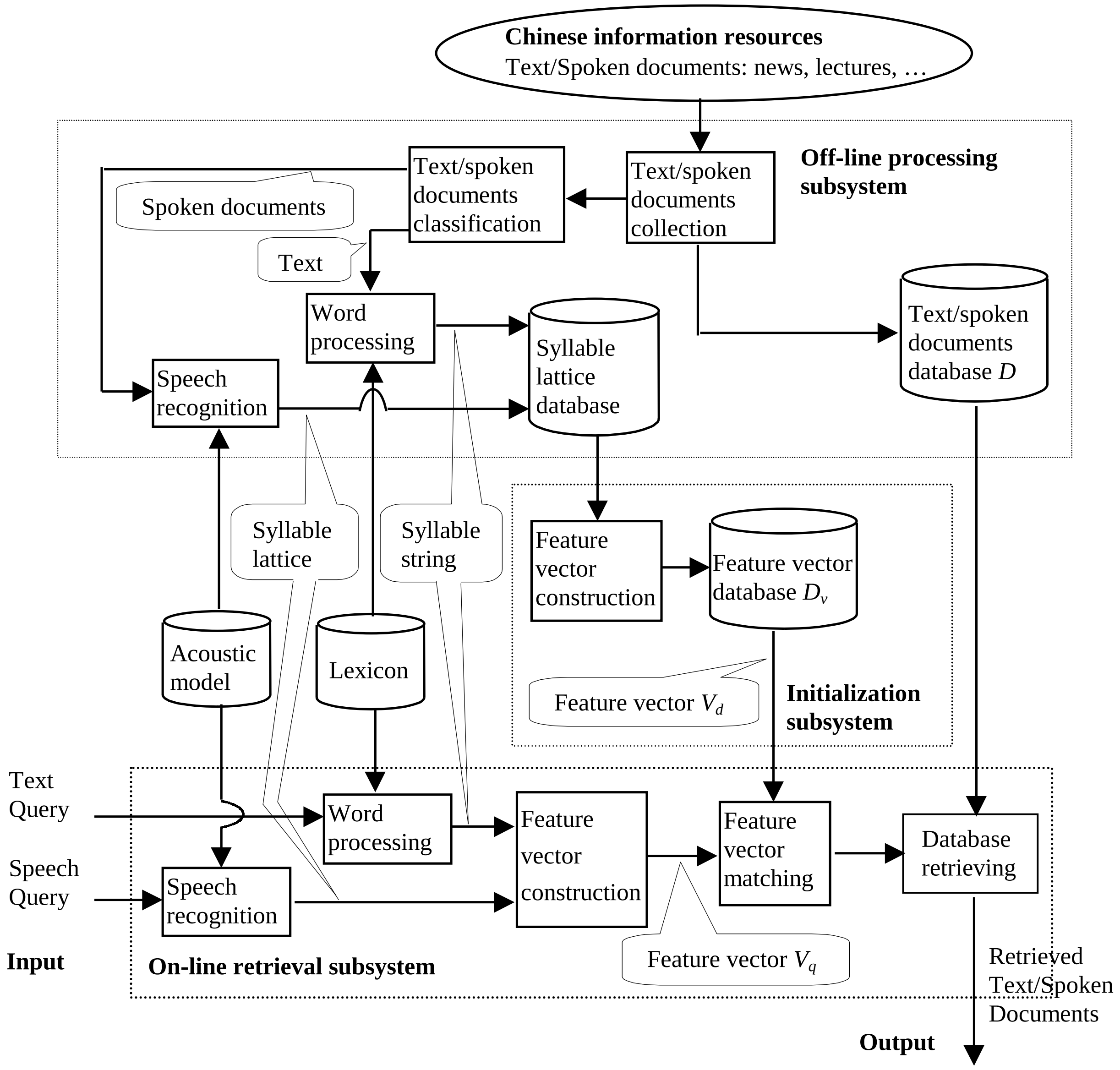
Chinese information resources
Text/Spoken documents: news, lectures, …
Off-line processing subsystem
Text/spoken documents collection
Text/spoken documents classification
Spoken documents
Text
Word processing
Speech recognition
Syllable lattice database
Text/spoken documents database D
Syllable lattice
Syllable string
Acoustic model
Lexicon
Feature vector construction
Feature vector database D_v
Feature vector V_d
Initialization subsystem
Text Query
Speech Query
Word processing
Speech recognition
Feature vector construction
Feature vector matching
Database retrieving
Feature vector V_q
Input
On-line retrieval subsystem
Retrieved Text/Spoken Documents
Output

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

database including 5.3 hours of speech for phonetically balanced sentences and isolated words produced roughly by 80 male and 40 female speakers. Also, to deal with the silence segments in the spoken documents or speech queries, a single state HMM is used to represent the silence.

Based on the acoustic models mentioned above, the syllable recognition process for the spoken documents is described as follows [9]. First, all possible beginning frames of syllables in the spoken document will be obtained by some spectral information. Corresponding to a beginning frame, the possible ending frames for the syllable can be found using the estimated minimum and maximum duration of a syllable. Given all these possible beginning frames and their corresponding ending frames, with a dynamic programming approach several most possible syllable candidates with their acoustic scores can be found for each segment which may include a syllable. A syllable lattice thus can be easily obtained by backtracking through the entire utterance. Exactly the same procedures can be applied to speech queries to generate the corresponding syllable lattices. Of course, speech recognition is performed off-line in advance for spoken documents, but on-line in real time for speech queries.

4. Feature Vectors

Before describing the details of the retrieving process, here we will introduce the feature vectors used for similarity measure in advance. For each document *d* in the database *D*, through searching the syllable lattice, all acoustic scores of single syllables and adjacent syllable pairs in the syllable lattice can be extracted to form the feature vector V_d ,

$$V_d = (as(s_1),...,as(s_i),...,as(s_{416}), \\ as(s_1,s_1),...,as(s_i,s_j),...,as(s_{416},s_{416})) \quad (1)$$

where $as(s_i)$ is the acoustic score of the syllable s_i , and $as(s_i,s_j)$ is the acoustic score of the syllable pair (s_i,s_j) . For a text document, we use the frequency counts instead of the acoustic score information to form the feature vector. The feature vector constructing procedures can be performed off-line on all documents in the database *D* to form a feature vector database D_v , which will be the target database to be physically retrieved.

In general, the syllable information is quite rough while the syllable pair information is more precise. However, the syllable information is very important in the task of Chinese text/spoken document retrieval. In Mandarin Chinese, many proper names, which are frequently used as the key information for retrieval, are very often 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, it's obvious that their adjacent syllable pairs do not match with each other at all, though they actually represent the same organization. That is, one can't obtain the documents contain "中央研究院" using the query "中研院", and vice versa. Furthermore, it's also obvious that the word order in Mandarin Chinese is very often not unique, 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". They both contain exactly the same 5 syllables, but only 2 of the 4 adjacent syllable pairs are the same, i.e., "院長" and "遠哲". In addition, when the speech recognition errors such as deletions, insertions, and substitutions occur in speech queries or spoken documents, sometimes the desired documents can't be retrieved 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, thus the documents containing "中研院" won't be retrieved. On the other hand, if the same recognition errors occur in a spoken document which contains "中研院", then it won't be retrieved by the query "中研院", either. This is why we incorporate both syllable information and syllable pair information in the feature vector.

While regarding a query, the same feature vector constructing procedures must be performed on-line to construct the feature vector V_q right after the input query is entered. On the other hand, to reduce the ambiguity caused by the irrelevant words that are very frequently contained in quasi-natural-language queries, such as "我想要找…(I would like to find…)", "有沒有關於…(Is there anything about…)", and so on, the inverse document frequency [10] 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 retrieving problem is now a searching process to retrieve the document d^* in the target database D_v which is most related to the query. This searching process can be formulated as follows:

$$d^* \equiv \arg\max_{d \in D_v} Sim(d,q) \quad (2)$$

where *Sim(d,q)* is a similarity measure between a document *d* and the query *q*. Here, a Cosine measure is used to estimate the similarity:

$$Sim(d,q) = \cos(V_d, V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|} \quad (3)$$

In this way, the larger the $Sim(d,q)$ value is, the more the document *d* is relevant to the query *q*. Documents with larger $Sim(d,q)$ values thus will be selected and ranked as the results.

6. Experiments and Discussions

This section will present several experiments to show the feasibility of the above approaches. The database used in the following experiments is introduced 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 in Taiwan area in 1997. The spoken documents were produced by 5 male speakers. Each speaker was asked to read 100 of the 500 text documents. On average, each document contains about 100 characters (i.e., 100 syllables), while their individual length ranges from 44 to 269 characters. On the other hand, 160 speech queries produced by 4 male speakers were used for testing, and they were further manually transcribed into text queries. 80 of the queries are simple queries, each contains only one key phrase for some news item without any irrelevant words. An example key phrase is "亞太經合會", which is a frequently used 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), those documents relevant to each query were identified manually in advance for performance evaluation purposes. The number of documents relevant to each query ranges from 1 to 20.

For each of the spoken documents and the speech queries, 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 construct the whole syllable lattice database used in the following experiments. Note that, in general, the text and spoken documents of a practical text/spoken database could contain different contents. However, the same materials used here are actually very helpful in evaluating the degree of difficulties among four different categories of retrieval problems.

6.2 Chinese Text/Spoken Document Retrieval Using Simple Queries

The first experiment was tested to show the performance of retrieving Chinese text/spoken documents using simple text/speech queries. Figure 2 shows the results in the recall-precision graph [11], where TQ, TD, SQ, and SD represents the text queries, text documents, speech queries, and spoken documents respectively, and thus the curve marked by “TQ/TD” shows 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 can achieve the best performance, and using speech queries to retrieve spoken documents is the most difficult task because the recognition errors might occur in both speech queries and spoken documents. Compared with the results of using text queries to retrieve text documents, large performance degradation was found for other three categories of retrieval that speech recognition is necessary to be applied in either queries or documents, or both. Using speech queries to retrieve spoken documents is obviously the worst case. This is reasonable since in this case, both the information to be retrieved and the input queries are in form of speech instead of text, thus with unknown variability on both sides. But in the other two categories, relatively precise information is available on one side. Also, it can be found that using speech queries to retrieve text documents gives slightly better performance than using text queries to retrieve spoken documents. In fact, the non-interpolated average precision rates [11] of these two categories of retrieval are 0.63 and 0.58 respectively. This is because it is much more difficult to recognize automatically the longer spoken documents than the relatively short input speech queries. Another interesting observation is that the two curves for text document retrieval are much more flat than the curves for spoken document retrieval. For spoken document retrieval, only relatively rough information is available on the spoken document side because of the recognition errors, more documents are therefore needed to include all relevant documents. That is, to obtain the high recall rate, the cost is a significant precision reduction.

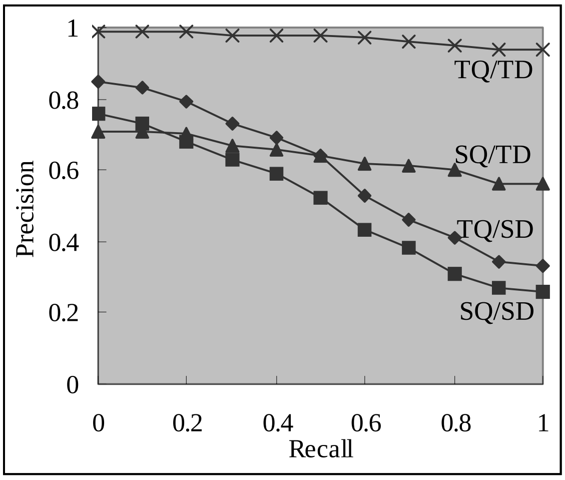

Figure 2: The performance of Chinese text/spoken document retrieval using simple queries.

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

This experiment was performed to evaluate the performance of retrieval using quasi-natural-language queries. The results in the recall-precision graph are shown in Figure 3. In general, very similar trend as we have discussed in the previous experiment for simple queries can be observed again in this figure, except that the performance difference between using speech queries to retrieve text documents and using text queries to retrieve spoken documents is less obvious here. In fact, their non-interpolated average precision rates are 0.53 and 0.52 respectively. This is because automatic recognition of long quasi-natural-language queries is no longer a simple problem as automatic recognition of short simple queries. Furthermore, the results for using simple queries and quasi-natural-language queries in four categories of retrieval are shown together for comparison in Figure 4 in non-interpolated average precision. For each category of retrieval, simple queries of course give better results than quasi-natural-language queries, but the performance difference is in fact not very significant, especially for the category of using text queries to retrieve text documents. These results indicate that the inverse document frequency information used here can more or less reduce the

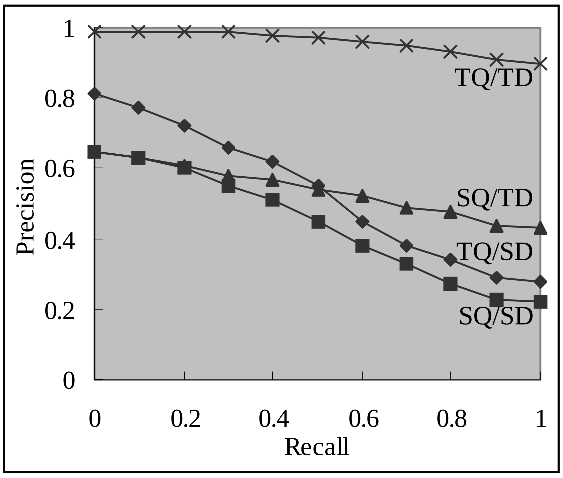

Figure 3: The performance of Chinese text/spoken document retrieval using quasi-natural-language queries.

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

7. Concluding Remarks and Future Work

This paper presents the initial results of a long-term research project towards speech retrieval. Due to the popularity of the Internet and multimedia, this research area has become very important. In this paper, we present a syllable-based approach for retrieving Chinese text/spoken documents using both text and speech queries. The experimental results reported here are not necessarily satisfactory, and it is believed that there are still many problems unsolved. However, these preliminary results at least show the very good potential in this direction.

Currently, we are trying to evaluate our approach using a large real-world database consisting of news broadcasts. Automatic recognition of such spoken materials of course is a very challenging problem. Furthermore, the relevance feedback techniques, such as query expansion schemes and so on, that are widely used in conventional text information retrieval are currently under study to improve the retrieving performance.

References

[1] Lee-feng 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.

[2] K. Spärck 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.

[3] Glavitsch, U. and Schäuble, P. "A System for Retrieving Speech Documents", *ACM SIGIR Conference on R&D in Information Retrieval*, pp. 168-176, 1992.

[4] Julian Kupiec, Don Kimber and Vijay Balasubramanian, "Speech-Based Retrieval Using Semantic Co-Occurrence Filtering", *The Human Knowledge Technology Workshop*, pp. 373-377, 1994.

[5] Bo-Ren Bai, Lee-Feng Chien, and Lin-Shan Lee, "Very-Large-Vocabulary Mandarin Voice Message File Retrieval Using Speech Queries", *ICSLP96*, pp. 1950-1953.

[6] Kenney Ng and Victor Zue, "Subword Unit Representations for Spoken Document Retrieval", *EUROSPEECH97*, pp. 1607-1610.

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

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

[9] Hsin-min Wang, Lin-shan Lee, et al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 2, pp. 195-200, March 1997.

[10] G. Salton, "Introduction to Modern Information Retrieval", NY, McGraw-Hill, 1983.

[11] Donna Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)", available at "http://trec.nist.gov/pubs/trec4/overview.ps".

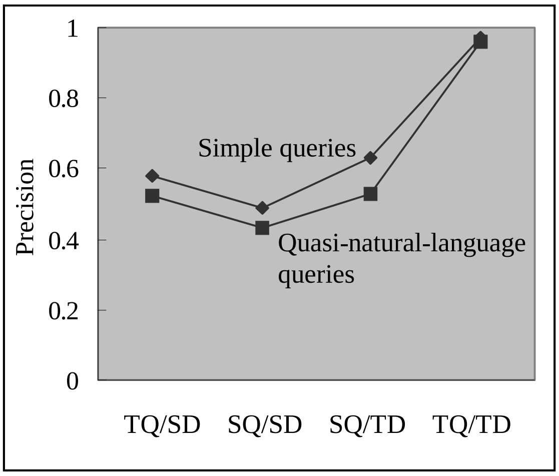

Figure 4: The comparison between using simple queries and quasi-natural-language queries.