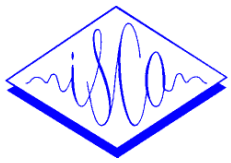

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A SYLLABLE-BASED VERY-LARGE-VOCABULARY VOICE RETRIEVAL SYSTEM FOR CHINESE DATABASES WITH TEXTUAL ATTRIBUTES

Sung-Chien Lin[1] ,Lee-Feng Chien[2], Keh-Jiann Chen[2], Lin-Shan Lee[1,2]

[1] Dept. of Computer Science and Information Engineering, National Taiwan University
[2] Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
e-mail address: lsc@speech.ee.ntu.edu.tw

ABSTRACT

In this paper a syllable-based voice retrieval approach for Chinese textual databases retrieval is presented. The presented approach can reduce most of difficulties of Chinese voice retrieval and easily integrate with the continuous speech recognition technology of the Mandarin dictation machine, Golden Mandarin (III). The experimental results show that the presented approach is easy to implement and systems based on it can allow users to retrieve Chinese textual databases using spoken queries and unconstrained vocabulary. Although the proposed approach is statistics-based and has some restrictions in linguistic analysis, the achieved results are very encouraging and have shown its feasibility in creating practical applications which demand the recognition ability of very large vocabulary.

I. INTRODUCTION

Use of speech recognition technology in information retrieval for databases provides users with a convenient computer interface environment . For Chinese language, because the language is not alphabetic and the input of Chinese characters into computers is still a difficult and unsolved problem, voice retrieval of Chinese databases, especially of Chinese textual databases, becomes apparently an important application area of Mandarin speech recognition. For developing a spoken language processing system which can handle very large vocabulary, a specially designed voice retrieval approach is presented in this paper. The presented approach has successfully integrated with the continuous speech recognition technology of the Mandarin dictation machine, Golden Mandarin (III)[3] and has proved that can allow users to retrieve Chinese textual databases using spoken queries and unconstrained vocabulary.

However, to develop a textual database retrieval system which can accept voice-input queries is very difficult. In addition to the possible errors of speech recognition resulted from the complexity of large vocabulary, to pursue such an approach there are many difficulties to be overcome, for example, the input query could be expressed with complicated semantics, the sorts of mentioned attribute values could be different, the order of attribute values in the queries could be very free, and some irrelevant terms could be introduced in the queries. To deal with the above difficulties, the presented approach is a kind of syllable-based and statistics-oriented approaches.

Considering the special mono-syllabic characteristic of Chinese language, the presented approach is completely syllable-based to reduce the difficulties of speech recognition. This approach perform a phonetic level matching between the recognized syllable string of the input query and the transcribed syllabic version of all textual values in the database. Although phonetic information is less exact in semantics, such a syllable-based approach still conceives many advantages including, not only the higher accuracy and the faster recognition speed in Mandarin syllable recognition rather than those in character recognition, but also the reduction of the difficulties caused by large vocabulary size and proper nouns.

On the other hand, in avoidance with using complicated natural language understanding technology, the presented approach is statistics-oriented rather than linguistic-oriented[4]. This approach does provide a very realistic solution for creating practical applications. The presented approach adopts a relevance-ranking algorithm to replace the use of grammatical analysis in retrieving Chinese textual databases. For each input spoken query, it directly matches the query with each record in the database to obtain a relevance score by carefully considering the statistics similarity between the contents of the query and the examined record. As a consequence, the data records with higher relevance scores are taken as the results.

For realizing the performance of the presented approach, an experimental system in retrieving a realistic Chinese bibliographic database has been implemented and extensive tests were done. The experimental results show that the presented approach is easy to implement and allows users to retrieve Chinese textual databases using spoken queries and unconstrained vocabulary. Although the proposed approach has some restrictions in linguistic analysis, the achieved results are very encouraging and have shown its feasibility in creating practical applications which demand the recognition ability of very large vocabulary, for examples, e-mail retrieval , legal document retrieval and news retrieval [5].

In the rest of the paper, an overview of the proposed approach is first introduced in Section II, and then the method of relevance estimation is described in Section III. Moreover, some experimental results and concluding remarks are given in Section IV.

II. THE PROPOSED APPROACH FOR VOICE RETRIEVAL

Before describing the voice retrieval approach in detail, the problem to be discussed will be first defined. In our work the problem of voice retrieval can be formalized as the searching process to retrieve the record r^* in the target database D which is the most related to the given spoken query Q.

$$r^* \overset{def}{=} \arg\max_{r \in D} R(r,Q) \quad ...(1)$$

where r represents any possible record in database D, $R(r, Q)$ is the estimated relevance value between r and Q, and r^* is the request record which has the highest relevance value. Based on this definition, the specially designed syllable-based approach is then developed and described below.

The Syllable-based Approach

Because every attribute value in the textual database is possible to be used in the spoken query for retrieval, the number of words to be recognized is very large. The recognition of the spoken queries are thus difficult, especially when the input query contains some meaningful proper nouns. Fortunately, such a difficulty can be effectively reduced using the syllable-based approach, due to the special mono-syllabic characteristic of Chinese language. In Chinese language, every word is composed of from one to several characters, and all of these characters are mono-syllabic. Since there are more than 10,000 Chinese characters but the total number of Mandarin syllables is only 1,345, the recognition of Mandarin syllables is believed reliable than that of characters. For reducing the recognition difficulties, the presented approach is completely syllable-based; that is, the entire processing of voice retrieval is performed on the phonetic level rather than the character-level. Initially, all of the records in the searching textual database are transcribed to their syllabic versions in advance. For each input spoken query it has to be recognized as the corresponding syllable string before retrieval. The estimation of relevance scores is, therefore, performed with a syllable-based scheme. For this reason, the above Equation (1) can be extended as follows:

$$\begin{aligned} r^* &\overset{def}{=} \arg\max_{r \in D} R(r,Q) \\ &\overset{def}{=} \arg\max_{r \in D} R(r_S,Q_S) \end{aligned} \quad .... (2)$$

where Q_S and r_S represent the recognized syllable string of the input spoken query Q and transcribed syllabic version of record r, respectively.

The Statistics-oriented Relevance Ranking

As mentioned in Section I, it has diverse uncertainties in the spoken queries. To alleviate these difficulties and avoid using complicated natural language understanding technology, the presented approach is statistics-based. The presented approach adopts a relevance-ranking algorithm in retrieving Chinese textual databases without analyzing the detailed semantics of each input spoken query but directly matches the query with each of the records in the database to obtain a relevance score. Such a relevance score is estimated by carefully considering the statistics similarity between the contents of the query and the examined data record. As a consequence, the data records which have higher relevance scores are taken as the results. Hence, the problem of voice retrieval for Chinese textual databases can be further extended as follows.

$$\begin{aligned} r^* &\overset{def}{=} \arg\max_{r \in D} R(r_S,Q_S) \\ &\overset{def}{=} \arg\max_{r \in D} R((a_1^r,a_2^r,\ldots,a_m^r),Q_S) \\ &\overset{def}{=} \arg\max_{r \in D} \sum_{i=1}^{m} Sim(a_i^r,Q_S) \end{aligned} \quad ...(3)$$

where a_i^r represents the i-th attribute value of record r (it has been transcribed as its syllabic version) and $Sim(a_i^r, Q_S)$ is the phonetic similarity between a_i^r and Q_S. At this stage, the entire approach has been formed and the detailed procedure listed in Fig. 1.

Firstly, in the first step, the input spoken query Q will be transcribed into the syllable string Q_S by the speech recognition subsystem. Then, for each record r in database D it will determine a relevance score $R(r_S, Q_S)$ in the second step. The estimation of $R(r_S, Q_S)$ is the core technique of the proposed approach. At first, for each attribute value a_i^r in r it will have a similarity value $Sim(a_i^r, Q_S)$ to represent its relevance to the query (the details will be described in the next section). Then, for record r its $R(r_S, Q_S)$ can be further obtained by summing up all of the composed $Sim(a_i^r, Q_S)$. In the third step, r^* which has the highest relevance score will be found and taken as the result.

III. THE ESTIMATION OF RELEVANCE SCORES

In this section a two-pass method to estimate $Sim(a_i^r, Q_S)$ will be described in detail. In the first pass, a fast approximate string matching is performed to extract a chunk which is the most similar to the examined attribute value. Then, in the second pass, the relevance score is calculated by estimating the statistic similarity between the extracted chunk and the attribute value according to a cosine measure.

The Approximate Syllable String Matching

As mentioned above, the goal of the first pass processing is to find a chunk q_i^r in the recognized syllable string of the query, $Q_S=Q_{S1}....Q_{SL}$, which is most similar to the examined attribute value $a_i^r = a_{i1}^r \ldots a_{il}^r$, where a_{ij}^r and Q_{Sk} represent the j-th syllable and the k-th syllable in a_i^r and Q_S, respectively. Because there may exist some differences between a_i^r and Q_S with forms of insertion, deletion, and substitution of syllables, in the proposed approach, an approximate syllable string matching algorithm is adopted with that a_i^r is then taken as the searching pattern and Q_S the text to be searched.

The adopted algorithm consists of four processing steps. Initially, the syllable string of the query is scanned and an array **A** with size L (the length of the syllable string of Q_S) is established to record if the composed syllables and syllable

pairs of Q_S appear in a_i^r. Every element in array **A** is then assigned a three-value weight after matching Q_S with a_i^r. For example, occurs

$$\mathbf{A}[j] = \begin{cases} 0 & \text{if } Q_{Sj} \text{ does not occur in } a_i^r \\ 1 & \text{if } Q_{Sj} \text{ occurs in } a_i^r \text{ but } Q_{Sj-1} \text{ not} \\ 2 & \text{if } Q_{Sj-1}Q_{Sj} \text{ occur in } a_i^r \end{cases}$$

If syllable Q_S does not occur in a_i^r, this syllable seem to be an error of speech recognition or an irrelevant character to a_i^r. So that, **A**[*j*] is given a weight 0. Otherwise, Q_S occur in a_i^r and we give **A**[*j*] a weight 1. Besides, considering most of keywords for retrieval are poly-syllabic in Chinese, if syllable pair $Q_{Sj-1}Q_{Sj}$ occurs in a_i^r, the weight of **A**[*j*] is then changed to be 2.

After weights of all elements in array **A** have been determined, in the chunk segmentation step it searches all of the syllable chunks in Q_S of which all corresponding component weights in **A** are greater than zero, *i.e.*,

$$\{Q_{SI} \cdots Q_{SF} | \mathbf{A}[j] > 0, \text{for } I \le j \le F, \text{and } \mathbf{A}[I-1] = 0, \text{and } \mathbf{A}[F+1] = 0\}$$

Then, considering the possibility of insertions or substitutions of syllable in Q_S, it concatenates two adjacent syllable chunks if they are separated with only one irrelevant syllable in the chunk concatenation step. Finally, in the chunk identification step it finds out the most similar syllable chunk by summing up all weights in the chunks and selecting the chunk with the maximal sum.

For illustration, an example shown in Fig. 2 is given, where the string "abcde" represents the syllable string of the examined attribute value a_i^r and "pqabsdetubcdv" the recognized syllable string of query Q_S. In the first step, the value of each element **A**[*j*] in array **A** is determined as (0,0,1,2,0,1,2,0,0,1,2,2,0). From **A**, it can find three chunks "ab", "de", and "bcd" in the second step. After performing concatenation in the third step, "absde" and "bcd" are the two remaining chunks. Because the weight of chunk "absde" is 6 that is greater than 5 of "bcd", chunk "absde" is then selected as the result of this example. In fact, the above processing is able to deal with the uncertainties mentioned in Section I. For example, if the examined attribute value exactly matches the request of the query, the corresponding part in the query could be therefore extracted, even if the query contains some irrelevant terms and more than one attribute values requested.

Relevance Estimation Using Cosine Measure

To estimate the relevance value between the input query and the examined attribute value, in the second pass the syllable strings of both the attribute value and the matched chunk are transformed into the forms of their feature vectors. For each component in the feature vectors, it indicates the frequency count of the syllable occurring in the attribute value or the chunk but weighted by an IDF value $idf(S_i)$. For example, $\mathbf{u}_{a_i^r}$ is the feature vector assigned to the attribute value a_i^r,

$$\mathbf{u}_{a_i^r} = (f(s_1) \times idf(s_1), \ldots, f(s_i) \times idf(s_i), \ldots, f(s_{1345}) \times idf(s_{1345}))$$

where $f(S_i)$ is the frequency count for syllable S_i in the attribute value a_i^r. The IDF value of syllable S_i, $idf(S_i)$, is defined as

$$idf(S_i) \stackrel{def}{=} \log(N / N_{S_i})$$

where N is the number of total attribute values in the database and N_{S_i} the number of attribute value containing syllable S_i. The IDF value assigns a high weight to syllables which are encountered in only a small number of attribute values in the database. It is supposed that rare syllables have high discrimination values and the occurrence of such a syllable in both an attribute value and a syllable chunk is a good sign that the attribute value is related to the query. The feature vector $\mathbf{u}_{q_i^r}$ is similarly defined. The estimation of relevance value is then computed based on the following "cosine measure" of these two feature vectors.

$$Sim(a_i^r, Q_S) \stackrel{def}{=} \cos(\mathbf{u}_{a_i^r}, \mathbf{u}_{q_i^r}) = \frac{\mathbf{u}_{a_i^r} \bullet \mathbf{u}_{q_i^r}}{|\mathbf{u}_{a_i^r}||\mathbf{u}_{q_i^r}|}$$

The larger value of cosine means that the attribute value is more related to the input query.

IV. EXPERIMENTAL RESULTS AND CONCLUDING REMARKS

To show the feasibility of the proposed approach, a voice retrieval system for Chinese bibliographic database retrieval has been implemented. This system successfully integrates the technology of Mandarin speech recognition and the proposed approach. The organization of the system is shown in Fig. 3, where it can be found that the system is composed of three subsystems: the data extraction subsystem, the speech recognition subsystem and the information retrieval subsystem. The speech recognition subsystem is actually the Golden Mandarin (III) and the information retrieval subsystem is designed according to the proposed approach. Meanwhile, the data extraction subsystem is served to extract the parameters from the databases for the applications of both speech recognition and information retrieval. Currently, the target database consists of about 30,000 bibliographic records extracted from a public library. This voice retrieval system allows users to search for each of the records using a spoken query. The query content is restricted in the attributes of the book title, the author name and the publisher. The query may contain some irrelevant terms and more than one of the mentioned attribute values. Besides the order of the attribute values could be free. For realizing the performance of the system, extensive tests were done. Below, some of the test results are shown.

The Test of Simple Query

In the first test, we randomly selected hundreds of book titles from the database and formulated each of them as a simple query for test (just single book title and no irrelevant words included). Meanwhile, for realizing the affectedness of the rate of recognition accuracy, the speech recognition results of test queries were also mixed with some noisy syllables. The performance was then evaluated by the hit rate for the retrieved records from top 1 to top 5. The test results are shown in Table. 1. It shows that the proposed approach can achieve good performance when the simple query with reasonable recognition errors.

The Test of Quasi-Natural-Language Query

The second test demonstrates the feasibility of using quasi-natural-language query to retrieve Chinese textual databases. Two hundred of quasi-natural language queries which contain one or more attribute values and some adjective words were formed. The hit rates for the retrieved top one records are listed in Table 2, where it shows that the proposed approach can also achieve good performance, even if the input is a quasi-natural language query and have some recognition errors included. Besides, it can be observed that the achieved retrieval performance is proportional to the number of mentioned attribute values in the queries.

Concluding Remarks

In this paper a syllable-based, statistics-oriented approach for voice retrieval of Chinese textual databases has been presented. The presented approach can reduce most of difficulties of Chinese voice retrieval, such as the unsolved problems of Chinese proper nouns, the errors of speech recognition, irrelevant words in input queries, and the uncertainties of ordering and positions of attribute values in input queries. Such an approach can easily integrate with the continuous speech recognition technology of the Mandarin dictation machine, Golden Mandarin (III). Systems based on the presented approach has been proved that can allow users to retrieve Chinese textual databases using spoken queries and unconstrained vocabulary. Although the proposed approach has some restrictions in dealing with complicated queries, the achieved results are very encouraging and have shown its feasibility in creating practical applications which demand the recognition ability of very large vocabulary.

V. REFERENCE

[1] C. Weinstein, "Demonstrations and Applications of Spoken Language Technology: Highlights and Perspectives from the 1993 ARPA Spoken Language Technology and Applications Day", ICASSP94, Vol. I, pp. 345-348, South Australia, 1994.

[2] V. Zue, et. al., "PEGASUS: A Dialogue Interface for On-line Air Travel Planning", Speech Communication, Vol. 15, pp. 331-340, 1994

[3] H-m. Wang, L-s. Lee, et. al, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary", ICASSP95, Vol. I, pp. 61-64, Detroit, U.S.A., 1995.

[4] C-H. Lee, "Stochastic Modeling in Spoken System Design", Speech Communication, Vol. 15, pp. 311-322, 1994

[5] L-F. Chien, " Fast and Quasi-Natural Language Search for Gigabytes of Chinese Text", to appear in SIGIR-95, ,1995.

Procedure of the Proposed Voice Retrieval Approach

Input: (1) A spoken query Q
(2) A Chinese textual database D

Output: the most promising record r^*

Step 1. to recognize Q into the syllable string Q_S

Step 2. for each record r in D to estimate the relevance score $R(r_S, Q_S)$

Step 2.1. for each attribute value a_i^r in r to estimate relevance score $Sim(a_i^r, Q_S)$

Step 2.2. to obtain $R(r_S, Q_S)$ by summing up all of the composed $Sim(a_i^r, Q_S)$

Step 3. to find r^* which has the highest relevance score as the output

Fig. 1 The procedure of the proposed voice retrieval approach

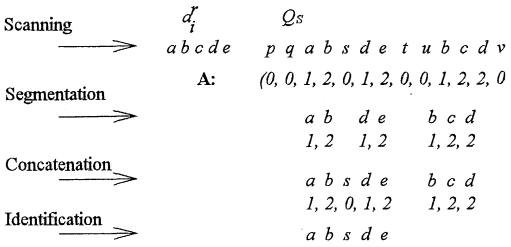

Fig. 2 An example to show the processing steps of the approximate syllable string matching

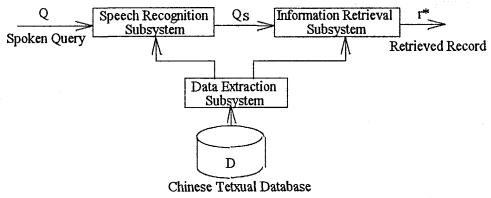

Fig 3. The organization of the voice retrieval system for Chinese bibliographic database retrieval

Hit Rate	A	B	C	D	E	F
Top 1 Record	.94	.90	.85	.80	.74	.56
Top 2 Records	.97	.94	.88	.85	.81	.62
Top 3 Records	.98	.95	.89	.85	.85	.65
Top 4 Records	.98	.96	.90	.88	.85	.65
Top 5 Records	.98	.97	.90	.88	.87	.67

Table. 1. The test results of hit rates with simple queries. Column **A**, **B**, **C**, **D**, **E** and **F** represent query with 0%, 5%, 10%, 15%, 20% and 30% errors, respectively.

Hit Rate	A	B	C	D	E	F
query with 1 Attr. Val.	.94	.90	.85	.80	.74	.56
query with 2 Attr. Val.	.99	.98	.97	.95	.91	.85
query with 3 Attr. Val.	1	1	1	.99	.95	.96

Table. 2. The test results of hit rates with quasi-natural language queries.

Column **A**, **B**, **C**, **D**, **E** and **F** represent query with 0%, 5%, 10%, 15%, 20% and 30% errors, respectively.