EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

INTELLIGENT RETRIEVAL OF VERY LARGE CHINESE DICTIONARIES WITH SPEECH QUERIES

Sung-Chien Lin[1], Lee-Feng Chien[2], Ming-Chiuan Chen[2] ,Lin-Shan Lee[1,2], Ker-Jiann Chen[2]
[1]Dept. of Computer Science and Information Engineering, National Taiwan University
[2]Institute of information Science, Academia Sinica
Taipei, Taiwan, Republic of China
lsc@speech.ee.ntu.edu.tw

ABSTRACT

To retrieve a Chinese word from a Chinese dictionary, it needs the user to know exactly the first character of the desired word. Because there is more than 10,000 Chinese characters, this makes the Chinese dictionary relatively difficult to be used. To reduce the problem, this paper presents intelligent retrieval techniques for very large Chinese dictionaries with speech queries. The proposed techniques properly integrate the technologies of Mandarin speech recognition and Chinese information retrieval with a syllable-based approach utilizing the mono-syllabic structure of the language. Moreover, it is very nice to provide the function of retrieving all relevant word entries from the dictionaries using speech queries describing "general concepts" of the desired words. To achieve the challenging function, the techniques of relevance feedback are also included. Based on these techniques, a retrieval system was implemented successfully on a Pentium PC for a very large Chinese dictionary which includes 160,000 word entries and the total length of the lexical information under the word entries exceeds 20,000,000 words.

I. INTRODUCTION

This paper presents intelligent retrieval techniques for very large Chinese dictionaries with speech queries. Because Chinese language is not alphabetic and there exist a very large number of Chinese characters, all word entries are ranked by the first character of the words in a Chinese dictionary and the users who want to retrieve a word entry in a Chinese dictionary should know the first character of the desired word. These cause the retrieval in a Chinese dictionary becoming a very difficult problems, even from the new electronic Chinese dictionaries. Using the proposed techniques of speech retrieval, users can easily retrieve the relevant word entries from Chinese dictionaries with unconstrained speech queries matching the syllabic structure or even describing the concepts of the desired word entries.

Due to the very unique character/morpheme mapping in Chinese language, the wording structure is very flexible in Chinese with unlimited number of compound words, word variants, abbreviation, etc. This makes the speech retrieval very difficult. The proposed techniques presented in this paper successfully solve these problems by properly integrating Mandarin speech recognition technologies with Chinese information retrieval technologies with a syllable-based approach utilizing the mono-syllabic structure of the language. Syllable and syllable-pair based statistical feature structures were used for representing the word entries, lexical information and the queries, and a dictionary-specific language model was used for linguistic decoding. All matching processes were performed on the syllable level. All word entries and their lexical information in the dictionary with the syllabic structure closed to the input speech queries can thus be retrieved efficiently.

A more challenging function is to try to retrieve all relevant word entries simply using unconstrained speech queries describing some "general concepts" of the desired words. To alleviate such a problem, the techniques of relevance feedback were also included. This is the lexical information of retrieved word entries is used to expand the retrieval subjects of the original query, such that word entries which are conceptually relevant but do not have similar syllable structure with the query are also able to be retrieved successfully. The relevance feedback techniques are also used to improve the performance of the dictionary-specific language model for the successive speech queries.

To test the feasibility of the proposed techniques, a working system on a Pentium PC has been implemented. The dictionary used in the system included 160,000 word entries and the total length of the lexical information under word entries exceeds 20,000,000 words. Accuracy on the order of 88-92% were achieved almost in real-time.

In Section II we will firstly present the problems and the difficult of Chinese dictionary retrieval. In Section III we give an overview of the proposed techniques for speech retrieval of Chinese dictionary. To retrieve word entries with their concepts the relevance feedback techniques is thus introduced in Section IV. In Section V we describe the working system for speech retrieval of Chinese dictionary based on the proposed techniques and conclude this paper with some general remarks.

II. PROBLEMS FOR CHINESE DICTIONARY RETRIEVAL

Every Chinese word is composed of from one to several characters. The traditional way to find a Chinese word entry from a dictionary is by the "head radical (部首)", "number of strokes(筆劃數)" or "phonetic structure" of the first character in the word. This requires the user to know exactly the first character of the word, either its shape or its pronunciation. Because Chinese language is not alphabetic and there exist at least 10,000 commonly used characters, this makes the Chinese dictionaries relatively difficult to be used. The new electronic dictionaries [1] didn’t help too much on this problem, because the input of Chinese characters into computers is very difficult due to the same reason. As a result, voice input for Chinese dictionary retrieval is highly attractive.

However, the number of commonly used word entries in an ordinary dictionary very often exceeds 100,000, but it is quite natural that a desired word is not within this huge number but some other similar words are. This is because every Chinese character is a morpheme with its own meaning, the wording structure is thus rather flexible with almost unlimited number of compound words, word variants, abbreviation, etc. Therefore a successful dictionary retrieval technology has to be able to handle such wording structure flexibility, which is difficult and special for Chinese language.

Furthermore, in many cases, the users try to obtain some help from the dictionary because they really have no definite ideas about the exact desired word entries, but instead only with some general concepts of the desired words, or sometimes they may like to retrieve all relevant word entries with similar concepts. For example, the users may wish to retrieve all word entries “*describing the beauty of a women*”, or all “*names of Nobel Prize laureates all over the world*”. Information for retrieving word entries with such “general concepts” does not exist very often in the explanations and example sentences under the word entries in most dictionaries, but correct retrieval of the word entries based on such speech input is not only very challenging but requires excellent technology in Chinese information retrieval, since the input queries can never match very well with such information.

III. CHINESE DICTIONARY RETRIEVAL USING SPEECH QUERIES

In this paper, very efficient and intelligent techniques for Chinese dictionary retrieval using speech queries having the above discussed fascinating functions were successfully developed. This is by properly integrating the technologies of Mandarin speech recognition and Chinese information retrieval with a syllable-based approach utilizing the special mono-syllabic structure of Chinese language [2,3,4]. All Chinese characters are monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1345, therefore very often each syllable represents quite several homonym characters each with different meaning, and the unlimited combinations of these 1345 syllables give the unlimited number of Chinese words with highly flexible structure. As a result, the syllables in Mandarin Chinese become very special linguistic units carrying plurality of linguistic information and each Chinese word can be seen as the combination of these 1345 Mandarin syllables. Therefore, in this approach the syllable-level statistical feature parameters are extracted from the speech queries and matched with the syllable-based feature structures of all word entries in the dictionary on syllable level. It turns out that by properly choosing the feature parameters and carefully designing the retrieval algorithm, the very challenging problem mentioned above can be solved very well.

The block diagram of the proposed technology is shown in Fig. 1. During the phase of database preparation, the feature analysis subsystem extracts the feature structures (based on syllable and syllable-pair statistics) of the word entries in the Mandarin dictionary. The feature structures are primarily based on syllable and syllable-pair statistics. Because syllables carry plurality of linguistic information and syllable-pairs can preserve the useful information of syllable ordering within words in Chinese language, the feature structure can be as linguistic constraints to guide the search of possible syllable strings for speech queries and also as the statistical indices assigned to queries and each word entry in the dictionary to estimate their relevance scores on syllable level. Therefore, using these syllable-based feature structures the feature analysis subsystem constructs a syllable-based, dictionary-specific language model and also a set of syllable-based feature vectors for the word entries.

When a user enters a speech query into the system, the speech recognition subsystem first transcribes the query into the most possible syllable string by acoustic recognition using the syllable acoustic models and linguistic decoding using the syllable-based, dictionary-specific language model. The parameters of the language model are constructed using the conditional probabilities of syllable bigram but are weighted by the inverse document factors (*idf* value), proposed in [5]. Since the language model is dictionary-specific, the transcribed syllable string is not only the most possible one but also relevant to retrieval with irrelevant syllables deleted.

The recognized syllable string of the query is then transformed into a feature vector in the information retrieval subsystem and compared with the syllable-based feature vectors of all word entries in the dictionary. Each component in a feature vector for the query and word entries is the frequency of a syllable or a syllable-pair in the query or the lexical information under the word entries and are weighted by the *idf* value of the syllable or the syllable-pair. The relevance scores of the word entries are estimated primarily based on the normalized inner product between the feature vector of the query and of all the word entries. As a consequence, those word entries with that the syllabic structures of them or their lexical information are close to the syllabic structures of the speech query can obtain high relevance scores. After all the word entries have be compared, the word entries with the highest relevance scores are selected as the results of retrieval.

In addition to being sent to user interface for display, the retrieval results can be fed back to the relevance feedback subsystem for extracting feature structures for further retrieval. Because the retrieved word entries with their lexical information contain very useful linguistic information to constrain the recognition of successive speech queries with the similar requested subjects, the syllable-based feature structures of these word entries and their lexical information can be extracted to adapt the syllable-based language model. Moreover, because the extracted feature structure can be used to expand the requested subjects of the query, they are also added to the original feature vector of the query to retrieve the word entries with related concepts to the query. For more declaration, the techniques of relevance feedback will be discussed in more details in the next section.

IV. RELEVANCE FEEDBACK TECHNIQUES

The relevance feedback techniques have been widely used in the area of information retrieval for several years [6]. The idea of relevance feedback is adding the terms in the retrieved relevant records to the original query to attempt expanding the requested subjects of the query for retrieving more relevant records in the database. In this section, the techniques used to improve the requested subjects of queries and the dictionary-specific language model are introduced.

For the difficult problem of retrieval with word concepts, there always exist some word entries which are conceptually relevant but do not have terms under the lexical information similar with the query, and therefore do not have similar syllable-based feature structure with the query. These word entries are difficult to be retrieved simply using the above approach. A possible solution is to adopt the techniques of relevance feedback to alleviate such problems, in other words, adding the retrieved results to the original query. A very nice nature of the dictionary retrieval is that terms in all the lexical information (i.e., synonyms, antonyms, word explanations and example sentences, etc.) under the word entries are undoubtedly relevant to the word entries and also the query. As a result, the lexical information of the word entries obtained in the first retrieval for a speech query can be extracted the syllabic feature structure and automatically added into the original query to expand the requested subject. The users can also select manually the most relevant parts of information obtained in the first retrieval to be added to the query. In this way, most desired word entries with relevant concepts to the query can be retrieved easily.

In some cases, users could retrieve word entries using longer queries with more complex requested subjects, such as using the speech query “a gas present in the air is necessary for all forms of life on Earth” to retrieve the desired word entry “oxygen”. The recognition of long queries is more difficult than that of a set of shorter queries with related concepts. Users may also prefer speak a set of shorter queries rather than a longer query. Such as the above example query can be replaced with a set of shorter queries, such as “a gas is present in the air” and “the gas is necessary for all forms of life on Earth”. It is obvious that the successive queries are related to at least one of the retrieved word entries and the lexical information of the word entries is helpful to improve the recognition of the successive queries. Therefore syllable-based feature structures of the feedback information are extracted to update the syllable-based language model, as also shown in Fig. 1.

V. EXPERIMENTAL SYSTEM AND CONCLUDING REMARKS

The above intelligent and efficient Chinese dictionary retrieval techniques have been successfully implemented as a working system on a Pentium PC under WIN 95 which operates almost in real-time. A Chinese dictionary containing about 160,000 word entries is used in the system, in which the total length of the lexical information under these word entries exceeds 20,000,000 words. The previously developed syllable recognition module [4] was used in the speech recognition subsystem, and the system can in fact accept unconstrained speech queries with all functions mentioned above. In the preliminary tests, the retrieval accuracy for queries carrying simply the desired word entry and utterances describing the desired word concept are 88% and 92% respectively. Fig. 2 is an example retrieval result for a query carrying only “general concepts”.

In summary, this paper presents intelligent retrieval techniques for very large Chinese dictionaries with speech queries for the difficult problems of Chinese dictionary retrieval. The proposed techniques is based on a syllable-based approach by properly integrate the technologies of Mandarin speech recognition and Chinese information retrieval and utilizing the mono-syllabic structure of the language. Moreover, a techniques of relevance feedback are included to retrieve all relevant word entries from the dictionaries using speech queries describing “general concepts” of the desired words. Based on these techniques, a retrieval system was implemented successfully on a Pentium PC for a Chinese dictionary with 160,000 word entries.

This paper presents not only the proposal of the retrieval techniques of very large Chinese textual databases via speech queries, but also the techniques of intelligent retrieval with the queries describing the concepts of the desired words. The success of intelligent retrieval is by properly using the feedback lexical information of the retrieved word entries to expand the request subjects of input queries and to update the dictionary-specific language model for the successive queries. Using the similar ideas, the relevance feedback techniques can be applied retrieval in other textual databases. The contents of retrieved records from the databases can be used to extract useful linguistic information for query expansion and language model adaptation to improve the retrieval performance. A database-specific thesaurus can also be automatically established using the contents of all records in the databases to expand the requested subjects of queries.

REFERENCES

[1] T. Yokoi, “The EDR Electronic Dictionary,” *Communications of the ACM*, Vol. 38, No. 11, pp.42-44, 1995.

[2] S-C. Lin, L-F. Chien, K-J. Chien, and L-S. Lee, “An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model,” ICASSP’96, pp. 287-290, Atlanta, USA, May, 1996.

[3] S-C. Lin, L-F. Chien, K-J. Chen, L-S. Lee, "A Syllable-Based Very-Large-Vocabulary Voice Retrieval Systems for Chinese Databases with Textual Attributes," EUROSPEECH’95, Vol. I, pp. 203-206, Sept., 1995.

[4] H-M. Wang, L-S. Lee, et. al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary, " ICASSP’95, Vol. I, pp. 61-64, Detroit, U.S.A., May, 1995.

[5] S-C. Lin, L-F. Chien, K-J. Chen, L-S. Lee, "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains," EUROSPEECH’95, Vol. I, pp. 203-206, Sept., 1995.

[6] D. Harman, “Relevance Feedback and Other Query Modification Techniques,” *Information Retrieval: Data Structures and Algorithms*, edited by W. Frakes and R. Baeza-Yates, Chapter 11, pp. 241-263, Prentice Hall, 1992.

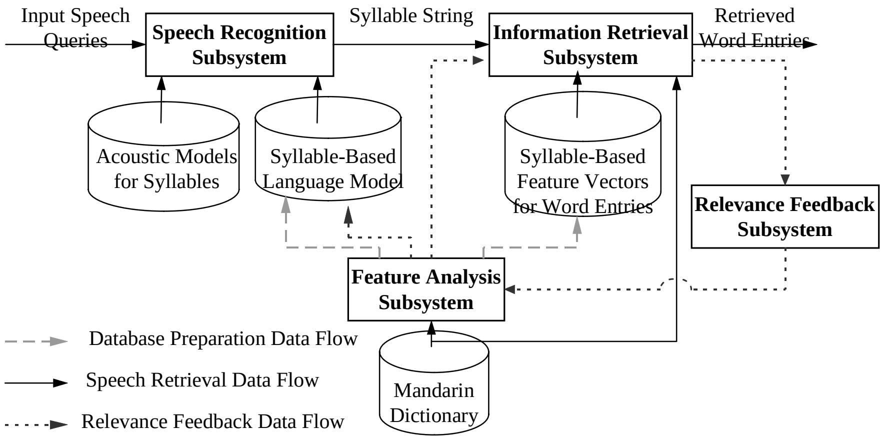
Input Speech Queries
Speech Recognition Subsystem
Syllable String
Information Retrieval Subsystem
Retrieved Word Entries
Acoustic Models for Syllables
Syllable-Based Language Model
Syllable-Based Feature Vectors for Word Entries
Relevance Feedback Subsystem
Feature Analysis Subsystem
Mandarin Dictionary
Database Preparation Data Flow
Speech Retrieval Data Flow
Relevance Feedback Data Flow

Fig. 1 Block Diagram of the Intelligent Chinese Dictionary Retrieval Techniques Using Speech Queries.

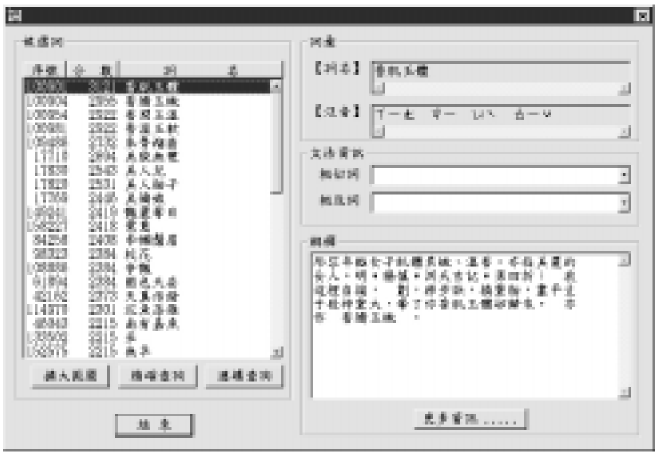

Fig. 2 An Example Retrieval Result for a Speech Query Carrying “General Concept”. The “General Concept” Entered is to “Find All Word Entries Describing the Beauty of a Woman”.