An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model

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
email-addr: lsc@speech.ee.ntu.edu.tw

ABSTRACT

This paper presents an accurate and efficient voice retrieval system for very-large-vocabulary Chinese textual databases with a specially-designed clustered language model. To reduce the problems resulted from the complexity of unconstrained speech-input queries for retrieval, the system is completely syllable-based in both speech recognition and database retrieval by properly utilizing the mono-syllabic structure of Chinese language. In addition, it partitions the records in the database into clusters and trains the clustered language model using the clustering results. The proposed clustered language model with its augmented search algorithm are very useful to improve accuracy and speed of the speech retrieval system. In the preliminary tests using an experimental database with about 30,000 bibliographical records, it was found that the present system can accept unconstrained speech-input queries and achieve very good performance.

I. INTRODUCTION

This paper presents an accurate and efficient voice retrieval system for very-large-vocabulary Chinese textual databases with a specially designed clustered language model. This system allows users to retrieve Chinese textual databases using unconstrained speech-input queries. Because Chinese language is not alphabetic and the input of Chinese characters into computers is still a very difficult and unsolved problem, the development of such a system becomes one of important and emergent research directions in the applications of Mandarin speech recognition [1].

For voice retrieval of Chinese textual databases, several difficult problems need to be alleviate[1], including the possible errors of speech recognition resulted from the complexity of very large vocabulary, the different sorts of mentioned attribute values, the free order of attribute values in the queries, and the irrelevant terms for the natural-language queries. As the size of database becomes very large, the retrieval speed also is a very critical factor for developing such a voice retrieval system.

To reduce the above problems, in this present research a syllable-based voice retrieval system is developed by properly utilizing the mono-syllabic structure of Chinese language and successfully integrating the technologies of continuous Mandarin speech recognition technologies [2] and Chinese information retrieval. Although there exist more than 10,000 commonly used Chinese characters, each character is mono-syllabic and the total number of phonetically allowed Mandarin syllables is only 1345. The combination of these mono-syllabic characters or 1345 syllables gives almost unlimited number of (at least 100,000 are commonly used) mono-syllabic or poly-syllabic Chinese words. This nice feature of Chinese language makes it possible to develop a syllable-based voice retrieval system for Chinese textual databases with very large vocabulary and unlimited domains [1]. In this way the present system is completely syllable-based in both speech recognition and database retrieval. When an unconstrained speech-input query is entered, it is firstly recognized into most possible syllable strings relevant for retrieval, with the syllable-based language model trained from the content of the database. The recognized syllable strings are then compared with records in the database to compute the relevance scores on syllable level. Finally, the records with higher scores are selected as the results.

For further reduction of possible errors and time consumption for retrieval, a specially-designed clustered language model (CLM) is proposed to provide more powerful linguistic constraints for speech recognition. With this language model and its augmented search algorithm, the results of speech recognition are not only the possible syllable strings for retrieval, but also the possible clusters of records containing relevant data in the database. Since the relevant clusters of records can be determined in advance, the syllable-level relevance estimation can take place only on records in the clusters instead of all records in the database.

In the preliminary tests using an experimental bibliographic database with about 30,000 bibliographic records grouped as 183 clusters, the hit rate for retrieved records is found to be more than 90% even if the syllable recognition accuracy is only on the order of 80%. Besides, about 93% of irrelevant records can be skipped and the retrieval speed can be significantly improved over 10 times, compared with conventional exhaustive search without using CLM.

In the following, the remaining parts of the paper are organized as follows. In section II, an overview of the complete voice retrieval system for Chinese textual databases is given. Three major subsystems of the system, including the clustered language modeling subsystem, the speech recognition subsystem, and the database retrieval subsystem, are separately described in the next three sections. Finally, section VI presents some experimental results of this system and concludes with some remarks.

II. THE VOICE RETRIEVAL SYSTEM FOR CHINESE TEXTUAL DATABASES

The block diagram of the present voice retrieval system for Chinese textual databases is shown in Fig. 1, which is composed

0-7803-3192-3/96 $5.00©1996 IEEE

of three subsystems. The clustered language modeling subsystem transcribes all attribute values in the database into the corresponding syllable strings, then partitions records in the databases into clusters based on the their syllable structure characteristics, and finally trains a database-specific clustered language model using the clustering results, when the databases are prepared. These processes are dealt with three modules in the subsystem, the syllable transcription module, the database clustering module, and the language model training module, respectively. When voice retrieval takes place, the speech recognition subsystem and the database retrieval subsystem cooperate to retrieve relevant records. The speech recognition subsystem is composed of a syllable recognition module and a syllable string searching module. The former is exactly the same acoustic recognition module previously developed for a Mandarin dictation machine [2], which can produce a lattice of possible syllable candidates for the speech-input query, and the latter module searches on the syllable lattice with the clustered language model and its augmented search algorithm. The results of the subsystem are top N possible syllable strings for the query information and the clusters containing relevant records of these syllable strings. The database retrieval subsystem finally compares these syllable strings with records in the determined clusters to estimate relevance scores using the relevance estimation method proposed in [1] and retrieves the records with higher scores.

III. CLUSTERED LANGUAGE MODELING SUBSYSTEM

The function of the clustered language modeling subsystem is to partition all records in the database into several clusters and to train a clustered language model. This subsystem comprises three modules: the syllable transcription module, the database clustering module, and the language model training module.

To facilitate voice retrieval in Chinese textual database, all attribute values of records in the database are transcribed into syllable strings in the syllable transcription module. For example, attribute values with the book title "國語語音辨認技術" ("Mandarin speech recognition technology") and the author name "王大明" ("Da-Ming Wang") of a record in a bibliographic database can be transcribed as the syllable string "guo2-iu3-iu3-yin1-bian4-ren4-ji4-shu4" and "wang1-da4-ming2". The method used in this module is based on a word identification algorithm [3] which matches the character strings of attribute values with a lexicon storing a large number of Chinese words with their corresponding syllables. The success rate for word identification of the adopted algorithm is as high as 99.77%, which is accurate enough for the present research.

After all attribute values are transcribed into syllable strings, the database clustering module partitions records, which hold similar syllable structure characteristics, into several clusters. Firstly, a binary feature vector is assigned to each record in the database. Each bit in this vector indicates the presence of a given syllable or a given syllable pair in the attribute value of the record. With these binary feature vectors, the similarity between two records R_i and R_j based on their syllable structure characteristics can be defined as

$$S(R_i, R_j) \stackrel{def}{=} \sum_k (b_{ik} \cdot b_{jk}) \cdot \sigma_k \quad \text{.....(1)}$$

where b_{ik}, b_{jk} are the values of the kth component in the binary feature vectors for the records R_i and R_j, respectively. σ_k is a weighting factor or the significance parameter of the kth component in the feature vectors,

$$\sigma_k \stackrel{def}{=} [\sum_i b_{ik}]^{-1}$$

If a given syllable or syllable pair that occurs in both records R_i and R_j appears in almost every record, its presence doesn't bring too much information and should be de-emphasized. With this definition, the records with similar syllable structure characteristics will have higher value and they can be clustered together using a nearest-neighbor clustering algorithm [4].

The transcribed syllable strings for attribute values and the resulting clusters are then used to train CLM by the language model training module. CLM consists of a global sub-model and a set of cluster sub-models, and is primarily syllable-pair-based. Each parameter in every sub-model is a weighted frequency count of a syllable pair obtained from the syllable strings of attribute values either in the complete database (for the global sub-model) or the specific cluster (for the cluster sub-model), which are defined as

$$wf_G(S_i,S_j) \stackrel{def}{=} f_G(S_i,S_j) \times idf(S_i,S_j)$$

$$wf_C(S_i,S_j) \stackrel{def}{=} f_C(S_i,S_j) \times idf(S_i,S_j)$$

where $wf_G(S_i,S_j)$ and $wf_C(S_i,S_j)$ are called as weighted frequency counts of the syllable pair (S_i,S_j) in the global sub-model and the cluster sub-model for a particular cluster C, and $f_G(S_i,S_j)$ and $f_C(S_i,S_j)$ are the frequency counts which (S_i,S_j) appears in all transcribed attribute values in the complete database and the cluster C, respectively. The weighting value $idf(S_i,S_j)$ indicates the significance of the information of the syllable pair (S_i,S_j) when (S_i,S_j) comes into the input query. This weighting value is defined as

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

where N and $N(S_i,S_j)$ are the total number of attribute values of the records in the database and the number of the transcribed attribute values containing (S_i,S_j). For example, if the syllable pair (S_i,S_j) is contained in all of the attribute values in the database, the information of (S_i,S_j) is less important. This CLM is then used in the speech recognition subsystem to find the most relevant syllable strings and the clusters of relevant records for the query.

VI. SPEECH RECOGNITION SUBSYSTEM

When a speech query is entered, the speech recognition subsystem is to recognize the query into top N syllable strings and the most promising clusters which may contain relevant records. First, the syllable recognition module produces the top n candidates for each syllable based on the syllable recognition technologies presented in [2], and these candidates are used to construct a syllable lattice as in Fig 2(a), where each syllable S_{ij} represents the top j candidate for the ith syllable in the query.

In order to delete the irrelevant syllables for retrieval and tolerate the possible speech recognition errors, the syllable string searching module in the subsystem firstly transforms the syllable lattice in Fig. 2(a) into a relevant syllable-pair lattice as shown in Fig. 2 (b). On the latter lattice, each node indicates the connection situation of a pair of syllables relevant for retrieval such as (S_{12}, S_{23}) under the condition that the two component syllables S_{12}, S_{23} not only connect in the syllable lattice of Fig. 2(a), but also appear adjacently in some attribute values in the database as indicated by CLM. In this way many irrelevant syllable pairs/syllables will be deleted. An arc which represents an entity of a syllable is then used to connect two nodes on this lattice if the corresponding ending syllable of the previous node is exactly the corresponding beginning syllable of the next node. For example, arc S_{23} indicates the connection between node (S_{12},S_{23}) and node (S_{23},S_{31}). Also, a class of special nodes called dummy nodes are added to each syllable pair position to take care of the irrelevant syllable pairs or recognition errors for retrieval. A dummy node can connect any node on its previous and next stages in the syllable-pair lattice.

A two-pass search algorithm is used in searching for N syllable strings on the relevant syllable-pair lattice and the most promising record clusters of these syllable strings. In the forward pass, we assign the scores for traveling through nodes and arcs using the weighted frequencies in the global sub-model and the acoustic scores obtained from the syllable recognition module. On each node, we compute the largest traveling score via all possible paths from the beginning of the lattice to this node using the modified Viterbi search algorithm [6].

$$TS(x) \stackrel{def}{=} \max_{y \in Prev(x)} [TS(y) + AS(Syl(y,x))] + wf_G(SP(x)) \quad \text{.....(2)}$$

where Prev(x) is the set of the nodes in the previous stage which are connected with x in the syllable-pair lattice, y is a node belonging to the set, and TS(x)and TS(y) are the largest traveling scores from the beginning of the lattice to x and y. Meanwhile, Syl(y,x) is the corresponding syllable of the arc connecting node x and node y, and AS(Syl(y,x)) is the acoustic score of the syllable. Moreover, SP(x) represents the corresponding syllable pair of node x, and wf_G(SP(x)) is its weighted frequency in the global sub-model. These accumulated scores are then used to guide the backward search.

In backward search to find N most possible syllable strings and the relevant clusters of records, we set two constraints. First, every path being searched through has to be composed of the dummy nodes or the nodes with the corresponding syllable pairs belonging to the same record cluster. Second, for each cluster only the most possible searching path will be reserved. The proposed backward search is a kind of tree search algorithm with using the heuristics of CLM and acoustic scores obtained from the syllable recognition module. Using the algorithm, the score of a searching path P, from the ending of the lattice to be extended to the node x, is defined as, f(P,x)=g(P,x)+h(P,x), where h(P,x) is the accumulated score computed in the forward pass and stored in the node x, that is =TS(x), and g(P,x) is the sum of the scores for traveling through the nodes and arcs on P using the weighted frequencies in the cluster sub-model and acoustic scores.

$$g(P,x) \stackrel{def}{=} \sum_{y \in P} wf_C(SP(y)) + \sum_{z \in P} AS(Syl(z)) + AS(Syl(x,P)) + wf_C(SP(x)) \quad \text{.....(3)}$$

where y and z are a node and an arc on P, respectively, and wf_C(SP(y)) is the weighted frequency of the syllable pair corresponding to node y in the cluster sub-model for the cluster C. Syl(x,P) is the corresponding syllable of arc connecting to x and P. Meanwhile,. With CLM and the tree search algorithm, N paths traveling through the relevant syllable-pair lattice can be found, and the syllable strings of these paths together with their corresponding clusters of records can also be obtained.

V. DATABASE RETRIEVAL SUBSYSTEM

The database retrieval subsystem compares the recognized syllable strings of the input queries with attribute values of records to estimate the relevance scores, and selects the records with higher scores as results. The relevance estimation method used in the subsystem is also a two-pass approach, proposed in [1]. The only difference of the present system is that the relevance estimation is performed merely with the records in the determined clusters rather than all records in the total database. The relevance estimation method is summarized below and the more detailed description can be referred to [1]. In the first pass of the method, a chunk on the recognized syllable string of the query, of which the syllable structure characteristics are most close to those of the examined attribute value, is identified. Afterward, in the second pass, both the syllable string of the examined attribute value and the identified syllable chunk are transcribed in syllable-based feature vectors, and the relevance score of the attribute value to the query is estimated as the cosine value between these two feature vectors. Finally, the relevance score of a record is estimated as the sum of the relevance scores of all of its attribute values.

VI. EXPERIMENTS AND CONCLUDING REMARKS

Experiments

In the preliminary tests of the present voice retrieval system, a Chinese bibliographic database containing about 30,000 bibliographic records was used as the experimental database. After database clustering, we obtained 183 clusters of records. 3 speakers provided 900 queries; that is, 300 different queries were uttered in continuous Mandarin by each speaker. These queries contain request information of from one to three attribute

values, including titles of books, names of authors, and names of publishers; also a few irrelevant words are involved in the queries. The final performance of the system is evaluated by the hit rate of the top 1 selected records.

The experimental results are shown in Table 1. In this table, each column demonstrates the experimental results of a speaker. The second row shows the results of syllable recognition; they are the accuracy rates of the top 1 syllable candidates in its first sub-row and the inclusion rates of the top 10 syllable candidates in the second sub-row, respectively. The third row shows the hit rates of top 1 selected records, which are results of exhaustive examination of all records without the clustered language model and taken as the baseline for test. The fourth row shows the experimental results with the clustered language model. It consists of three sub-rows, i.e. the inclusion rates of top 10 determined clusters, the ratios of the number of skipped examined records to the number of records in the total database, and the hit rates of the top 1 selected records.

From this table, it can be observed that even if the syllable recognition accuracy is only on the order of 80%, the hit rate of the top 1 selected records can achieve more than 90%. CLM and its augmented search algorithm can improve efficiency and accuracy of voice retrieval. A huge number of records can be skipped to be examined and the hit rates of the retrieved records increase substantially.

Concluding Remarks

This paper presents an accurate and efficient voice retrieval system for Chinese textual databases with very large vocabulary and unlimited domains. This system can allow users to retrieve large Chinese textual databases using unconstrained speech-input queries. With the special feature of the mono-syllabic structure in Chinese language, the system is completely syllable-based in both speech recognition and database retrieval to alleviate several difficult problems resulted from the complexity of unconstrained speech-input queries. In addition, this system presents the clustered language model (CLM) trained from the content of the databases as the useful linguistic constraints in speech recognition. With CLM and its augmented search algorithm, the results of speech recognition in the system are not only the possible syllable strings relevant for retrieval but also the clusters containing possible records, such that it can avoid examine all records in database retrieval. This makes voice retrieval more accurate and more efficient.

Based on these features, the system had successfully integrated the technologies of continuous Mandarin speech recognition and Chinese information retrieval. Although the system can only retrieve records based on the statistical characteristics at the syllable level, it has proven its fitness on the application of Mandarin speech recognition for database information retrieval. Furthermore, it should be noted that the proposed clustered language model is also useful for future work on word and topic spotting [5,6] in Chinese language, to further extend the spoken language processing applications.

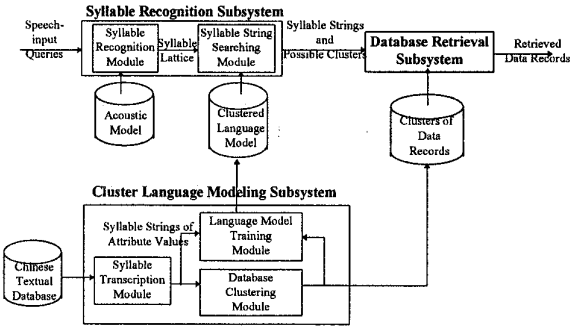

Fig. 1 The block diagram of the present voice retrieval system

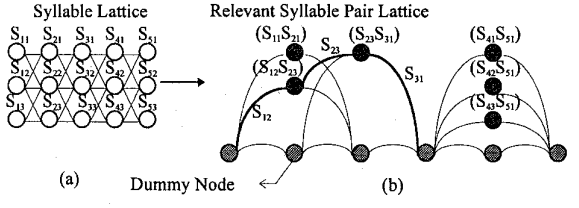

Fig. 2. The lattice in the speech recognition subsystem

		Speaker 1	Speaker 2	Speaker 3
Speech Recognition	Top 1 Syllable Accuracy Rate	85.41%	84.82%	78.14%
	Top 10 Syllables Inclusion Rate	98.49%	99.10%	95.16%
Baseline without CLM	Hit Rate of Top 1 Record	93.67%	96.00%	91.33%
the Present System with CLM	Inclusion Rate of Top 10 Clusters	99.33%	98.67%	96.33%
	Ratio of Skipped Records	93.25%	93.23%	93.36%
	Hit Rate of Top 1 Record	98.33%	97.00%	92.33%

Table 1. The results of the experimental system

Reference

[1] 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," EUROSPEECH95, Vol. I, pp. 203-206, Sept., 1995.

[2] 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., May, 1995.

[3] K-J. Chen and S-H. Liu, "Word Identification for Mandarin Chinese Sentences," Proc. of COLING-92, pp. 101-107, 1992.

[4] A. Jain and R. Dubes, Algorithms for Clustering Data, Prentice-Hall, New Jersey, 1988.

[5] E-F. Huang, H-C. Wang, and F-K. Soong, "A Fast Algorithm for Large Vocabulary Keyword Spotting Application," IEEE Trans. on Speech and Audio Processing, Vol. SAP-2, No. 3, pp. 446-449, 1994.

[6] L. Gillick, et. al., "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech," ICASSP93, Vol. II, pp. 471-474, Minneapolis, U.S.A., Apr. 1993.