Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching

Hsin-min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan 115, R.O.C.

E-mail: whm@iis.sinica.edu.tw

Abstract

Spoken document retrieval has been extensively studied in recent years because of its potential use in navigating large multi-media collections of the near future. We have been working on Mandarin spoken document retrieval. Considering the characteristic monosyllabic structure of the Chinese language, a syllable-based framework has been investigated, in which the similarity between the spoken documents and the speech queries was measured using the syllable and syllable pair information extracted from the syllable lattice based on the vector space models. We have recently developed a syllable-lattice-matching approach and a syllable lattice re-ranking scheme. By incorporating all the above methods, we achieved 0.25 absolute improvement in non-interpolated average precision over the baseline results.

1 Introduction

Spoken document retrieval has been extensively studied in recent years because of its potential use in navigating large multi-media collections of the near future [1-7]. With the rapidly growing audio and multi-media information on the Internet, a variety of exponentially increasing spoken documents such as the broadcast radio and television programs, video tapes, digital libraries and so on, are now being accumulated and made available via the Internet. But most of them are simply stored there, kind of difficult to be further reused for lack of efficient retrieval technologies. Development of the technology to retrieve speech information thus becomes essential and gets more and more important. Actually, this interest has been increased since the spoken document retrieval (SDR) track within the TREC-6 conference [8].

Speech recognition and information retrieval techniques enable automatic content-based indexing and efficient retrieval of spoken documents that are relevant to a user's query. That is, both the speech queries and the spoken documents must be transcribed into some kind of content features, such as phone strings or lattices, texts, keywords and so on using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can then be measured. Recognition errors usually degrade the effectiveness of a SDR system. Strategies against this problem at least include improving the speech recognition accuracy, selecting adequate indexing features, and developing retrieval methods that are more error-tolerant. In this paper, we focus on investigating the retrieval methods and also try to improve the speech recognition accuracy.

Researchers have addressed the problem of SDR in several different ways. For the keyword-based approach [2-3], a set of keywords for the spoken documents must be defined in advance, and whenever some keywords are extracted from the speech queries, the spoken documents with those or similar keywords can then be retrieved. This approach is efficient and cost-effective, and is very useful for retrieval of static databases with static queries, where the search words don't change frequently. However, usually it is not easy to define a set of adequate keywords for all the spoken documents to be retrieved unless we know the contents of all of them in advance. For the large-vocabulary-based approach, both the spoken documents and the speech queries are fully recognized into texts, thus many well-developed text retrieval techniques can be directly applied [6]. In the TREC SDR track, word-level transcripts of the spoken documents were provided to enable a participant without having to cope with speech recognition [8]. However, for such an approach, the out-of-vocabulary problem is an important issue. A large vocabulary speech recognizer needs a predefined lexicon for linguistic decoding, and some special words important for retrieval, such as proper nouns (e.g. personal names or organization names), exotic words, and domain specific terms (e.g. special terms for business news or sports news), may be simply outside of this predefined lexicon. This leads to the concept of making comparison on the level of subword units instead, or the subword-based approach [1,2,4,5,7]. Because it is much more easier to obtain all necessary subword units to cover all possible pronunciations of a given language, the out-of-vocabulary problem existing in the ever-growing speech information may be somehow handled by directly measuring the similarity between the spoken documents and the speech queries on the subword level instead of on the word level. Because in such approaches the subword units are never decoded into words, therefore the retrieval is never limited by any lexicon either. Such a subword-based approach also has the advantages of bypassing the complicated lexicon matching and linguistic decoding processes, in addition to avoiding the out-of-vocabulary problem.

Considering the monosyllabic structure of the Chinese language, the syllable-based approach has been found to be an attractive special case of the subword-based approaches for retrieving Chinese text [9] and speech [7] information using speech queries. In this approach, the subword unit selected is the syllable due to various considerations on the characteristics of the Chinese language. The similarity between the spoken documents and the speech queries is measured on the syllable level based on the vector-space models widely used in many traditional text information retrieval systems. The feature vector of each document or query contains the presence information, frequency counts, and acoustic recognition scores of all syllables and adjacent syllable pairs in the syllable lattice obtained by speech recognition. Here, using the syllable and syllable pair information extracted from the syllable lattice for Mandarin spoken document retrieval is a straightforward idea because more than half of the commonly used Chinese words are single character and bi-character words. However, the spoken document retrieval performance is obviously not satisfactory as compared to the upper-bound performance derived from text-

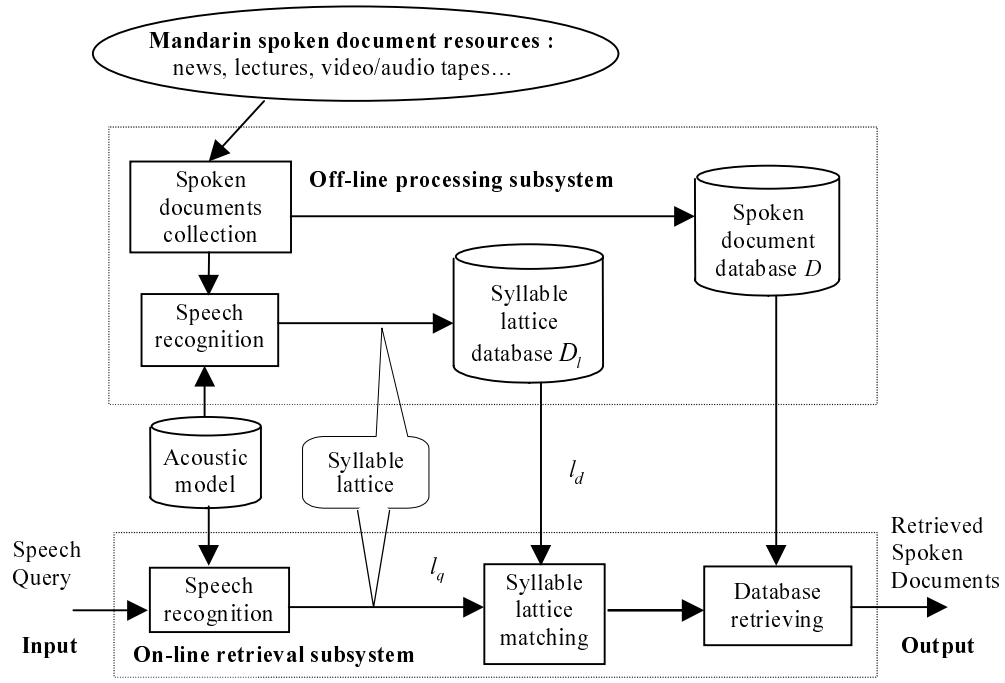


Figure 1: The overall architecture of the lattice-matching-based approach for retrieving Mandarin spoken documents using key-phrase speech queries.

based retrieval of transcripts of the spoken documents and the speech queries. The above vector-space-based approach was tested on the task of retrieving 500 spoken documents using 80 key-phrase speech queries [10], and the non-interpolated average precision [11] was 0.55 for speech retrieval and 0.97 for the upper-bound performance derived from text-based retrieval. Based on the lattice-matching-based approach, the non-interpolated average precision can be improved to 0.73. Since both the vector-space-based approach and the lattice-matching-based approach measure the similarity based on the same syllable lattice, they can be easily combined together. The non-interpolated average precision can be further improved to 0.76 by the combined approach. Furthermore, with the syllable lattice re-ranking scheme applied, the top-1 syllable recognition accuracy can be improved, and the non-interpolated average precision can be further improved to 0.80. Thus, we can achieve 0.25 absolute improvement in non-interpolated average precision over the baseline results.

The rest of this paper is organized as follows. The methodology of the lattice-matching-based approach is introduced in Section 2. The speech recognition and the retrieval method are presented in Sections 3 and 4 respectively. Section 5 briefly reviews the vector-space-based approach, and Section 6 introduces the combined approach. Finally, all experimental results are discussed in Section 7, and the concluding remarks are made in Section 8.

2 Methodology

The overall architecture of the lattice-matching-based approach for retrieving Mandarin spoken documents is shown in Figure 1. The whole system can be divided into two 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 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, for each collected spoken document, speech recognition with utterance verification techniques is first applied to generate a syllable lattice, including the acoustic recognition scores for all syllable candidates, and the syllable lattice is then added to the syllable lattice database D_l. In this way, the most time consuming speech recognition processes are performed off-line in advance, and all information necessary for retrieval is stored in the syllable lattice database D_l.

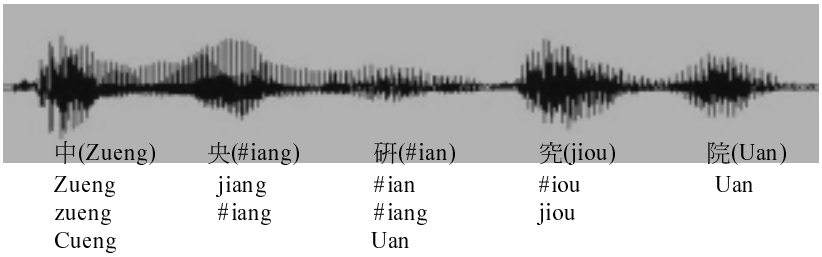
In the on-line retrieval subsystem, when a speech query is entered, speech recognition will first generate a syllable lattice for the speech query. Given the syllable lattice database D_l and the query syllable lattice l_q, the retrieving module then evaluates the similarity measures between l_q and all syllable lattices of the syllable lattice database D_l, and selects a set of documents with the highest similarity measures as the retrieval output.

3 Speech Recognition

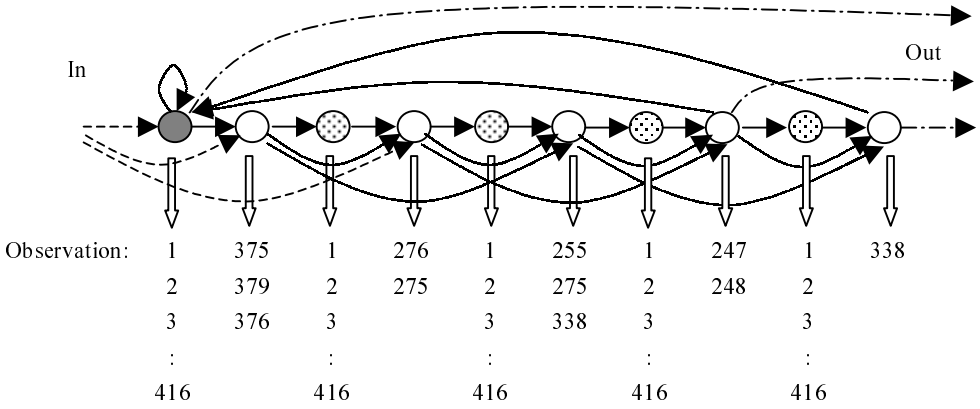
3.1 Feature Extraction

In our experiments, 13 mel-frequency cepstral coefficients (MFCC) were extracted from each speech frame, and these 13 MFCC coefficients and their first order and second order differences were combined together to form a 39-dimensional feature vector. In addition, to compensate the channel noise, utterance-based cepstral mean subtraction (CMS) is applied to all the training sentences, spoken documents and speech queries.

3.2 Acoustic Modeling



(a) syllable lattice



(b) DHMM representation of the syllable lattice

Figure 2: An example syllable lattice of a key-phrase speech query "中央研究院(Academia Sinica)", and the corresponding DHMM representation.

In Mandarin Chinese, there exist more than 10,000 commonly used characters. Each word is composed of from one to several characters, thus there exist almost unlimited number of words [12]. A nice feature of the language is that all characters are monosyllabic and the total number of phonologically allowed syllables is only 1,345. Furthermore, Mandarin Chinese is a tonal language, and the 1,345 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 large vocabulary Mandarin speech recognition as well as 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 the 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 112 context-dependent INITIAL's and 38 context-independent FINAL's specially considering the monosyllabic nature in Mandarin Chinese and the INITIAL/FINAL structure in Mandarin base syllables [13]. 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 was represented by a HMM with 3 states while each FINAL with 4 states. The state Gaussian mixture number ranged from 2 to 8. Therefore, every syllable unit was represented by a 7-state HMM. In addition, a 1-state HMM with 32 Gaussian mixtures was used to model the silence or non-speech part of the spoken documents and speech queries while another 1-state HMM with 32 Gaussian mixtures was used to model the breath effects. The above acoustic models were trained by the speech database with 5.3 hours of speech for phonetically balanced sentences and isolated words produced roughly by 120 speakers.

3.3 Syllable Lattice Construction

Based on the acoustic models mentioned above, the speech recognition processes for the spoken documents are described as follows: In the first pass, the speech recognizer performs the Viterbi search on the whole spoken documents and outputs the best syllable sequence and the corresponding syllable boundaries. In the second pass, based on the state likelihood scores calculated in the first pass search and the syllable boundaries of the best syllable sequence, the speech recognizer performs the Viterbi search on each utterance segment which may include a syllable and outputs several most possible syllable candidates with their acoustic recognition scores. Then, after the two-pass speech recognition processes, a syllable lattice can be constructed.

The acoustic recognition score, $\log p(O|s)$, for a certain syllable candidate s in the syllable lattice and the feature vector sequence O for a certain speech utterance segment is first normalized with respect to the duration of the observed speech segment, and then transformed into a range between 0 and 1 by a Sigmoid function,

$$\zeta(x) = \frac{1}{1+\exp\left(-\alpha\cdot(x-\beta)\right)} \tag{1}$$

where α and β are used to control the slope and the range of the Sigmoid function. Then, a simple utterance verification

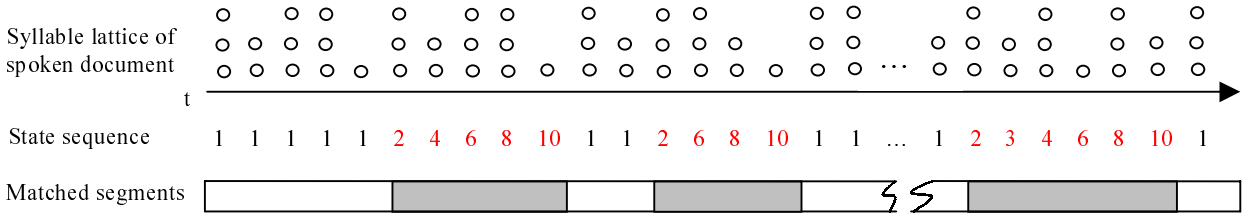


Figure 3: The syllable lattice matching process

scheme is used to filter out the syllable candidates with lower recognition scores. Initially, 20 syllable candidates are obtained for each syllable segment after speech recognition, while only those with the acoustic recognition scores larger than a threshold can be left after utterance verification. The depth of the syllable lattice thus can be adjusted by simply changing the threshold value, and a more compact syllable lattice can be obtained.

Exactly the same procedure can be applied to speech queries to generate the corresponding syllable lattices. Of course, the syllable lattice constructing processes are performed off-line in advance for spoken documents, but on-line in real time for speech queries.

3.4 Syllable Lattice Re-ranking

To improve the speech recognition accuracy is very important for spoken document retrieval. We have designed a syllable lattice re-ranking scheme to be applied in the syllable lattice before utterance verification. The re-ranking scheme performed the best-first search within the syllable lattice based on the acoustic recognition scores and the syllable bi-gram language models and selected the best syllable sequence. All the syllable candidates contained in the best syllable sequence were then re-ranked to the top-1 syllable candidate in the syllable lattice. After the re-ranking process, the utterance verification scheme was then applied. The syllable bi-gram language models were trained by a small text corpus consists of 25 million characters collected from Central News Agency (CNA) in 1999 from January to March.

4 Retrieval Method

Given the syllable lattice database D_l and a query syllable lattice l_q, the information retrieval problem is now a pattern matching process to identify the document d^* in the target database D_l whose syllable lattice l_{d^*} contains l_q. Here, we use a statistical method to estimate the possibility that a document d contains the key-phrase query q.

As shown in Figure 2, the syllable lattice of a key-phrase speech query can be represented as a discrete Hidden Markov Model (DHMM), $\lambda_q = (A, B, \pi)$ [14], where $A = \{a_{ij}\}$ is the state-transition probability distribution, $B = \{b_j(k)\}$ is the observation symbol probability distribution, and $\pi = \{\pi_i\}$ is the initial state distribution. The state number, N, equals to twice the length (i.e., the syllable number) of the speech query. The first state (the dark one, as shown in Figure 2) is the filler state that is used for decoding surrounding non-key-phrase part of the spoken document, thus its observations include all syllables and they all share the uniform observation probabilities, i.e., $b_1(k) = 1/416, 1 \le k \le 416$. The dotted states are also filler states and are used for handling the possible insertion errors of the spoken documents and deletion errors of the speech queries, thus their observations also include all syllables and they all share the uniform observation probabilities. On the other hand, the other states are the key-phrase states, which represent the corresponding syllable segments of the key-phrase query respectively, thus the observations of each state only include the syllable candidates, and their observation probabilities can be the acoustic recognition scores. That is, $b_{j\times 2}(k) = as(s_k)$, if s_k is one of the candidates of the j-th syllable of the speech query, otherwise, $b_{j\times 2}(k) = 0$. To handle the possible deletion errors of the spoken documents and insertion errors of the speech queries, the DHMM topology allows the search process to skip one key-phrase state each time. As a result, the distributions π and A can be easily derived according to the topology of the DHMM adopted here, as shown in Figure 2, e.g. $\pi_i = 1/3, i = 1,2,4$ while $\pi_i = 0, i = 3$, or $4 < i \le N$ because the entrance states include the first, second, and fourth states, and we would have $a_{ij} = 0$ for some (i,j) pairs. Furthermore, the exit states include the first state and the last two key-phrase states only.

The syllable lattice of a spoken document can be thought as an unknown sequence with multiple observations at each time index. Then, each spoken document is an unknown utterance and the speech query is the keyword model, while the retrieving processes should identify all the segments in the spoken document that are similar to the keyword and generate the accumulated scores of all the matched spoken segments as the similarity measure between the spoken document and the speech query. The diagram of the above process is shown in Figure 3. The first matched segment in Figure 3 represents two kinds of situations. The first one is a "perfect match", which means there exists a syllable sequence in this segment that is exactly the same as a syllable sequence contained in the query syllable lattice. The second one is that there are substitution errors between the query syllable lattice and the syllable lattice of this segment, but no insertion and deletion errors. The second and third matched segments in Figure 3 are obviously not exactly matched and there is a deletion error and an insertion error respectively. The details of the retrieval process are described in the following.

First of all, we can use the Viterbi search algorithm to find the best state sequence. The complete procedure is stated as follows [14]:

1.Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N \tag{2a}$$

$$\psi_1(i) = 0, \qquad 1 \le i \le N \tag{2b}$$

2.Recursion

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t), \quad 2\le t\le T, 1\le j\le N \tag{3a}$$

$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}], \quad 2\le t\le T, 1\le j\le N \tag{3b}$$

3. Termination

$$P^* = \max_{1\le i\le N}[\delta_T(i)] \tag{4a}$$

$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)] \tag{4b}$$

4. State sequence backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1 \tag{5}$$

where $O = (o_1 o_2 ... o_T)$ is the observation sequence, $\delta_t(i)$ is the best score (highest probability) along a single path, at time *t*, which accounts for the first *t* observations and ends in state *i*, and $Q^* = (q_1^* q_2^* ... q_T^*)$ is the best state sequence. In this approach, the estimation of $b_j(o_t)$ can be formulated as follows:

$$b_j(o_t) = \sum_{k=1}^{K} b_j(o_{tk}) \times as(o_{tk}) \tag{6}$$

where *K* is the number of syllable candidates contained in the syllable lattice of the spoken document at time index *t*, o_{tk} is the *k*-th syllable candidate contained in the syllable lattice of the spoken document at time index *t*, while $as(o_{tk})$ is the acoustic recognition score of the syllable candidate o_{tk}.

Then, based on the best state sequence, we can identify the matched spoken segments, as shown in Figure 3. Finally, we can estimate the similarity measure between a spoken document *d* and the speech query *q* using the following equation:

$$Sim(d,q) = \sum_{i=1}^{MSN} matched\_score(i) \tag{7}$$

where *MSN* is the number of matched spoken segments and $matched\_score(i)$ is defined as follows:

$$match\_score(i) = \sum_{t=t_i}^{t_i+D_i-1} b_{q_t^*}(o_t) \tag{8}$$

where t_i is the beginning time of the *i*-th matched spoken segment, and D_i is the duration of the *i*-th matched spoken segment. As a result, the documents with higher $Sim(d,q)$ will be selected and ranked as the retrieval results.

5 The Vector-space-based Approach

This section will briefly review our vector-space-based approach for retrieving Mandarin spoken documents. For each spoken document *d* in the database *D*, through searching the syllable lattice, all acoustic recognition scores of single syllables and adjacent syllable pairs in the syllable lattice were extracted to form the feature vector V_d,

$$\begin{aligned} V_d = {} & (as(s_1)\times idf(s_1), ..., as(s_{416})\times idf(s_{416}), \\ & as(s_1,s_1)\times idf(s_1,s_1), ..., as(s_i,s_j)\times idf(s_i,s_j), ..., \\ & as(s_{416},s_{416})\times idf(s_{416},s_{416})) \end{aligned} \tag{9}$$

where $as(s_i)$ is the acoustic recognition score of the syllable s_i, $as(s_i,s_j)$ is the acoustic recognition score of the syllable pair (s_i,s_j), while $idf(s_i)$ and $idf(s_i,s_j)$ are the inverse document frequency (IDF) of the syllable s_i and syllable pair (s_i,s_j) respectively. The feature vector constructing procedure was performed off-line on all documents in the database *D* to form a feature vector database D_v, which was the target database to be physically retrieved. While regarding a query, the similar feature vector constructing procedure was performed on-line to construct the feature vector V_q right after the input query is entered. Note that, in the following experiments, IDF was only applied in V_d but not in V_q. This is because in this paper, only the simple key-phrase queries were investigated. Given the feature vector database D_v and the query feature vector V_q, the Cosine measure was used to estimate the similarity between a document *d* and the query *q* [15]:

$$Sim(d,q) = \cos(V_d, V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|} \tag{10}$$

The documents with higher $Sim(d,q)$ were then selected and ranked as the retrieval results.

6 The Combined Approach

It should be noted that the computation requirement of the lattice-matching-based approach is much higher than that of the vector-space-based approach. Actually, in our experiments, we found that the search time of the lattice-matching-based approach is about 10 times of that of the vector-space-based approach, based on the same syllable lattice with the same verification threshold applied. Furthermore, usually the word order in Mandarin Chinese is relatively free, e.g. "President Lee Teng-Hui" can be expressed as "李登輝總統", "李總統登輝", "總統李登輝", and so on. Currently, the lattice-matching-based approach can't handle the word order problem, but the vector-space-based approach is obviously more robust to this phenomenon. Also, since both approaches measure the similarity based on the same syllable lattice, they can be very easily incorporated into a combined approach, which may achieve better retrieval performance. These lead to the idea of combining the similarity measures obtained by equations (7) and (10), and using a two-stage search strategy to shorten the search time. In the first stage, the vector-space-based approach is applied to filter out the non-relevant documents and select a set of potential documents. Then, in the second stage, the lattice-matching-based approach is applied to the potential documents only. Finally, these potential documents are re-ranked based on the summation of similarity measures obtained by two approaches and the final results are obtained. In the initial test, the lattice-matching-based approach was applied to 50 potential documents selected from 500 documents by the vector-space-based approach. In this case, the total search time is only about 2 (1+50/500×10) times of that for using the vector-space-based approach only.

7 Experiments

7.1 Test Settings

We have experimented on the methods described above using the test collection that consists of 500 Mandarin spoken documents of Chinese news. The collection consists of about 4 hours of speech materials, and was produced by 5 male speakers. The text materials are news articles published in Taiwan area in

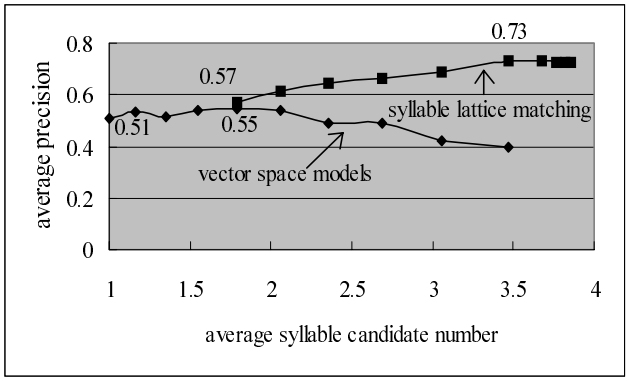


Figure 4: The non-interpolated average precision with respect to the average number of syllable candidates.

1997. It should be noted that the text materials and the text corpus for training the language models, as described in Section 3.4, are in different time ranges. On average, each spoken document contains about 100 characters (i.e., 100 syllables), while the individual length of each article ranges from 44 to 269 characters. The query set consists of 80 key-phrase speech queries produced by the other 4 male speakers. Each of these queries contains only a key-phrase for some news items. A typical 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)". These key-phrases were selected manually from the headlines of the original text materials. Each query contains 4.9 characters (or syllables) on average. For assessment of the retrieval performance, the relevant news articles for each query were selected manually. Each query has on average 5.9 relevant documents among the 500 documents in the database, with the exact number ranging from 1 to 20.

Gender-independent speaker-independent context-dependent INITIAL/FINAL HMM's as mentioned in Section 3 were used to recognize the syllables and construct the syllable lattices for both the spoken documents and the speech queries. The top-1 syllable recognition rates for the spoken documents and the speech queries are 54.70% and 59.07%, respectively. With the syllable lattice re-ranking scheme, as described in Section 3.4, applied, the top-1 syllable recognition rates for the spoken documents and the speech queries are improved to 71.04% and 79.07% respectively. The improvements in top-1 syllable recognition rates achieved by the simple re-ranking scheme are actually very significant. However, the improvements in retrieval performance are yet to be investigated.

7.2 Experimental Results

The first experiment was performed to compare the lattice-matching-based approach with the vector-space-based approach. The non-interpolated average precision with respect to the average number of syllable candidates is plotted in Figure 4. For the vector-space-based approach, it can be found that in general the performance becomes worse and worse when the number of syllable candidates increases, and the best average precision achieved is 0.55 when the average number of syllable candidates is only 1.79. When the number of syllable candidates is increased from 1 to n, the number of possible syllable pairs is increased from 1 to n×n. Although one of them may be correct and provide information regarding the desired documents, the other n×n-1 syllable pairs all include wrong syllables, and

	Vector space models			Syllable lattice matching	
Category	TQ/TD	SQ/SD	TQ/SD	SQ/SD	TQ/SD
Precision	0.97	0.55	0.63	0.73	0.77

Table 1: The non-interpolated average precision for different categories of retrieval.

therefore inevitably increase the degree of ambiguity. These explain why the performance degrades with increased number of syllable candidates using the vector-space-based approach. So good retrieving approaches should be able to make use of the increased correct syllables to achieve better results. For the lattice-matching-based approach, it can be found from Figure 4 that in general the performance becomes better when the number of syllable candidates increases, and the best average precision achieved is 0.73 when the average number of syllable candidates is 3.47. The lattice-matching-based approach produced 0.18 (0.73-0.55) absolute improvement in non-interpolated average precision, while the average number of syllable candidates in this case is almost double (3.47/1.79=1.94) the average number of syllable candidates used in the best case of the vector-space-based approach. It can also be found that, for the lattice-matching-based approach, the curve keeps relatively flat as the average number of syllable candidates further increases. These experimental results show that the lattice-matching-based approach is better than the vector-space-based approach in making use of the syllable lattice, and thus the retrieval performance is significantly improved.

The second experiment was performed to evaluate the retrieval performance using both speech queries and text queries. Here, the 80 text queries are the transcripts of the 80 key-phrase speech queries used in the above experiment. The texts were transcribed into syllable strings with 100% accuracy. These syllable strings can be thought as syllable lattices with only one candidate for each syllable and without any insertion or deletion errors. The results in non-interpolated average precision are summarized in Table 1, where SQ/SD and TQ/SD represent the results of spoken document retrieval using speech queries and text queries respectively. It is obvious from Table 1 that, for the lattice-matching-based approach, the performance of spoken document retrieval using speech queries (SQ/SD) is relatively close to that of spoken document retrieval using text queries (TQ/SD). On the other hand, for the vector-space-based approach, the performance for SQ/SD is relatively poor as compared to that for TQ/SD. The non-interpolated average precision for SQ/SD and TQ/SD are 0.73 and 0.77 for the lattice-matching-based approach, while 0.55 and 0.63 for the vector-space-based approach. The upper-bound non-interpolated average precision derived from text-based retrieval of transcripts of the spoken materials (TQ/TD) is also provided in Table 1 for reference. It is clear that though the lattice-matching-based approach can significantly improve the performance of Mandarin spoken document retrieval, the retrieval performance is still far behind that of text-based retrieval.

The third experiment was performed to evaluate the combined approach. The retrieval results in non-interpolated average precision are summarized in Table 2, in which the results obtained by simply combining the two approaches without using the two-stage search strategy are also shown together for reference. It can be found that the best non-interpolated average precision can be further improved to 0.76 and 0.78 for SQ/SD and TQ/SD respectively by the combined approach. However, in this case, the total search time is about 11 times of that for using

the vector-space-based approach only. With the two-stage search strategy applied, the non-interpolated average precision can be improved from 0.55 and 0.63 to 0.69 and 0.73 for SQ/SD and TQ/SD respectively. In this case, the total search time is only about twice as much as that for using the vector-space-based approach only.

The last experiment was performed to investigate the effectiveness of the syllable lattice re-ranking scheme. The retrieval results in non-interpolated average precision are summarized in Table 3. For the combined approach (CA), the non-interpolated average precision can be further improved to 0.80 and 0.83 for SQ/SD and TQ/SD respectively. Though the retrieval performance is obviously improved, the improvements are actually not as significant as the improvements in the top-1 syllable recognition rate. For the two-stage combined approach, the non-interpolated average precision can be improved to 0.79 and 0.81 for SQ/SD and TQ/SD respectively, which are in fact very close to the results of the combined approach. This is because, in this case, the vector-space-based approach recalled more desired documents in the 50 potential documents such that the lattice-matching-based approach could re-rank them to the top of the retrieved list.

	Combined Approach (CA)		Two-Stage CA	
Category	SQ/SD	TQ/SD	SQ/SD	TQ/SD
Precision	0.76	0.78	0.69	0.73

Table 2: The non-interpolated average precision for the combined approach with/without the two-stage search strategy applied.

	Combined Approach (CA)		Two-Stage CA	
Category	SQ/SD	TQ/SD	SQ/SD	TQ/SD
Precision	0.80	0.83	0.79	0.81

Table 3: The non-interpolated average precision for the combined approach with/without the two-stage search strategy applied (with the syllable lattice re-ranking scheme applied).

8 Conclusions

This paper first presented a lattice-matching-based approach for retrieving Mandarin spoken documents using key-phrase queries. This approach has been tested on simple key-phrase queries, and the experimental results show that the retrieval performance can be significantly improved with respect to the vector-space-based approach. However, whether this approach can be applied to Mandarin spoken document retrieval using natural language queries is yet to be further investigated. Since both the vector-space-based approach and the lattice-matching-based approach are based on the same speech recognition front-end, they can be easily incorporated into a combined approach and better retrieval performance can be achieved. Furthermore, to shorten the total search time, a two-stage search strategy has been preliminarily tested. Though the performance slightly degraded, the computation requirement was significantly reduced. We have also designed a syllable lattice re-ranking scheme to improve the top-1 syllable recognition rate using the syllable bi-gram information. By incorporating all the above methods, the best non-interpolated average precision achieved was 0.80 and 0.83 for spoken document retrieval using speech queries and text queries respectively, while the baseline results achieved by the vector-space-based approach were 0.55 and 0.63.

Acknowledgements

This work was partially supported by the Republic of China National Science Council under the grant No. NSC 88-2213-E-001-019. The author would like to thank Mr. Berlin Chen for providing the speech recognition front-end.

References

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

[2] D. A. James, "The Application of Classical Information Retrieval to Techniques to Spoken Documents", Ph.D. thesis, University of Cambridge, UK, 1995.

[3] K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, "Experiments on Spoken Document Retrieval", *Information Processing & Management*, 32(4), pp. 399-417, 1996.

[4] K. Ng and V. Zue, "Subword Unit Representations for Spoken Document Retrieval", In Proc. *European Conf. on Speech Communication and Technology*, 1997, pp. 1607-1610.

[5] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition" Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

[6] CMU Informedia Digital Video Library project "http://informedia.cs.cmu.edu/"

[7] B. R. Bai, L. F. Chien and L. S. Lee, "Very-Large-Vocabulary Mandarin Voice Message File Retrieval Using Speech Queries", In Proc. *Int. Conf. on Spoken Language Processing*, 1996, pp. 1950-1953.

[8] J. S. Garofolo, E. M. Voorhees, V. M. Stanford, and K. Spärck Jones, "TREC-6 1997 Spoken Document Retrieval Track Overview and Results", in Proc. *The sixth Text REtrieval Conference (TREC-6)*.

[9] S. C. Lin, L. F. Chien, K. J. Chen and L. S. Lee, "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains", In Proc. *European Conf. on Speech Communication and Technology*, 1995, pp. 1203-1206.

[10] B. R. Bai, B. Chen, and H. M. Wang, "Syllable-based Chinese Text/Spoken Document Retrieval Using Text/Speech Queries", In Proc. *Int. Conf. on Multimodal Interface*, 1999, pp. II46-II51.

[11] D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)", 1995, Available at "http://trec.nist.gov/pubs/trec4/overview.ps".

[12] L. S. Lee, "Voice dictation of Mandarin Chinese", *IEEE Signal Processing Magazine*, Vol. 14, No. 4, pp. 63-101, 1997.

[13] H. M. Wang, et al, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data" In *IEEE Trans. on Speech and Audio Processing*, 5(2), pp. 195-200, March 1997.

[14] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall International, Inc., 1993.

[15] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, NY, 1983.