A New Syllable-Based Approach for Retrieving Mandarin Spoken Documents Using Short Speech Queries

Hsin-min Wang

Institute of Information Science, Academia Sinica

Taipei, Taiwan 115, Republic of China

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

Abstract

Intelligent and efficient information retrieval techniques allowing easy access to huge amount and various types of information become highly desired and have been extensively studied in recent years. Considering the fast growth of audio resources and the characteristic monosyllabic structure of the Chinese language, a syllable-based framework of retrieving Mandarin spoken documents using speech queries has been investigated. This paper presents a new syllable-based approach that is based on matching the whole syllable lattice directly instead of using the syllable or syllable pair information extracted from the syllable lattice. The experimental results show that the retrieval performance can be significantly improved.

1. Introduction

The network technology and the Internet are creating a completely new information era. With the rapidly growing audio and multi-media information on the Internet, a variety of exponentially increasing spoken documents such as the broadcast radio, television programs, video tapes, digital libraries, courses and lectures 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 technology. Development of the technology to retrieve speech information thus becomes essential and gets more and more important. Recently, with the advances in speech recognition technology, proper integration of information retrieval and speech recognition has been considered by many researchers (Bai et al., 1996, 1999; CMU Informedia, Glavitsch and Schäuble, 1992; James, 1995; Jones et al., 1996; Ng and Zue, 1997; Wechsler, 1998). In any case, retrieval of speech information using speech queries directly is apparently the most natural, convenient and attractive, although the technology involved will be the most difficult as well. This is because in such cases both the information to be retrieved and the input queries are in form of voice instead of texts, thus

with unknown variabilities on both sides. For Chinese language, because the language is not alphabetic and the input of Chinese characters into computers has been a very difficult and unsolved problem even today, voice retrieval of speech information will be much more important and attractive for Mandarin Chinese than that for other languages.

Unlike the text information, the speech information can't be retrieved at all by directly comparing the input speech queries with the spoken documents. Therefore 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 or concepts and so on using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can then be measured. Thus, there can be at least the keyword-based, the large-vocabulary-based, and the subword-based approaches. For the keyword-based approach (James, 1995; Jones et al., 1996), one can define a set of keywords for the spoken documents 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 (CMU Informedia). However, for such an approach, the out-of-vocabulary problem is an important issue, since a large vocabulary speech recognizer needs a predefined lexicon for linguistic decoding, and some special words important for retrieval purposes, 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 (Bai et al., 1996, 1999; Glavitsch and Schäuble, 1992; James, 1995; Ng and Zue, 1997; Wechsler, 1998). 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 (Lin et al., 1995) and speech (Bai et al., 1996) 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. However, the spoken document retrieval performance is obviously not satisfactory as compared to the upper-bound performance derived from text-based retrieval of transcripts of the spoken materials. As previously reported in Bai's work (Bai et al., 1999), for simple key phrase queries, the non-interpolated average precision rates (Harman, 1995) are 0.97 and 0.54 for text retrieval and speech retrieval respectively. While, for quasi-natural-language queries, the non-interpolated average precision rate for speech retrieval is 0.43, which is even worse. Of course the serious performance degradation is due to speech recognition errors and the increased ambiguity comes from the syllable lattice itself. Although both the single syllable information and the syllable-pair information extracted from the syllable lattice maybe somehow robust to speech recognition errors, they are not precise enough and many wrong syllables and syllable pairs are also included in the feature vectors. So good retrieving approaches should be able to make use of the increased correct syllables contained in the syllable lattice to achieve better results. Accordingly, in this paper, we propose a new syllable-based approach that is based on matching the whole syllable lattice directly instead of using the syllable and syllable pair information extracted from the syllable lattice. This approach has been evaluated based on the task of Mandarin spoken document retrieval using short key phrase queries. The experimental results show that the retrieval performance can be significantly improved to 0.73, based on exactly the same task and the same speech recognition front-end previously used in Bai's evaluation.

In the following, the speech recognition process is first introduced in Section 2. Section 3 briefly reviews the previous syllable-based approach, and Section 4 presents the new syllable-based approach. Finally, all experimental results are discussed in Section 5, and the

concluding remarks are made in Section 6.

2. Syllable Lattice Construction

In Mandarin Chinese, there exists a total of 1,345 phonologically allowed tonal syllables, and these tonal syllables can be reduced to 416 base syllables and 5 tones. Base syllable recognition is thus believed to be the first key problem for 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 Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computation requirement in the recognition phase. Thus, context-dependent Initial/Final's (Wang et al., 1997) are widely used acoustic units for Mandarin speech recognition specially considering the monosyllabic nature in Mandarin Chinese and the Initial/Final structure in Mandarin base syllables. Initial is the initial consonant of the base syllable and Final is the vowel (or diphthong) part but including optional medial or nasal ending. Each Initial or Final is then represented by a left-to-right continuous HMM. To allow anyone to use the system naturally without training, the retrieval system is operated under the speaker-independent mode. That is, the speaker-independent context-dependent Initial/Final HMM's are used to recognize the syllables and construct the syllable lattices. These models are trained by a training speech database including 5.3 hours of speech for phonetically balanced sentences and isolated words produced roughly by 80 male and 40 female speakers. Also, to deal with the silence segments in the spoken documents or speech queries, a single state HMM is used to represent the silence.

Based on the acoustic models mentioned above, the speech recognition processes for the spoken documents are described as follows: First, the speech recognizer performs the Viterbi search (Rabiner and Juang, 1993) on the whole spoken documents and outputs the best syllable sequence and the corresponding syllable boundaries. Then, based on the state likelihood scores calculated in the Viterbi search and the syllable boundaries of the best syllable sequence, the speech recognizer performs the second Viterbi search on each utterance segment which may include a syllable and outputs several most possible syllable candidates with their acoustic recognition scores. After the above speech recognition processes, a syllable lattice can be easily constructed.

The acoustic recognition score, $\log p(O \mid s)$, 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 $\zeta(x)$,

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

where α and β are used to control the slope and the range of the sigmoid function. Here, a simple utterance verification scheme is used to filter out the syllable candidates with less possibilities. 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.

3. The Previous Approach

This section will briefly review the overall system architecture, the feature vector, and the retrieving process of our previous syllable-based approach for retrieval of Mandarin spoken documents (Bai et al., 1996, 1999). This approach is primarily based on the vector-space models widely used in many traditional text information retrieval systems.

3.1 Overall System Architecture

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

In the off-line processing subsystem, 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 process is performed off-line in advance, and all information necessary for retrieval is stored in the syllable lattice database D_l.

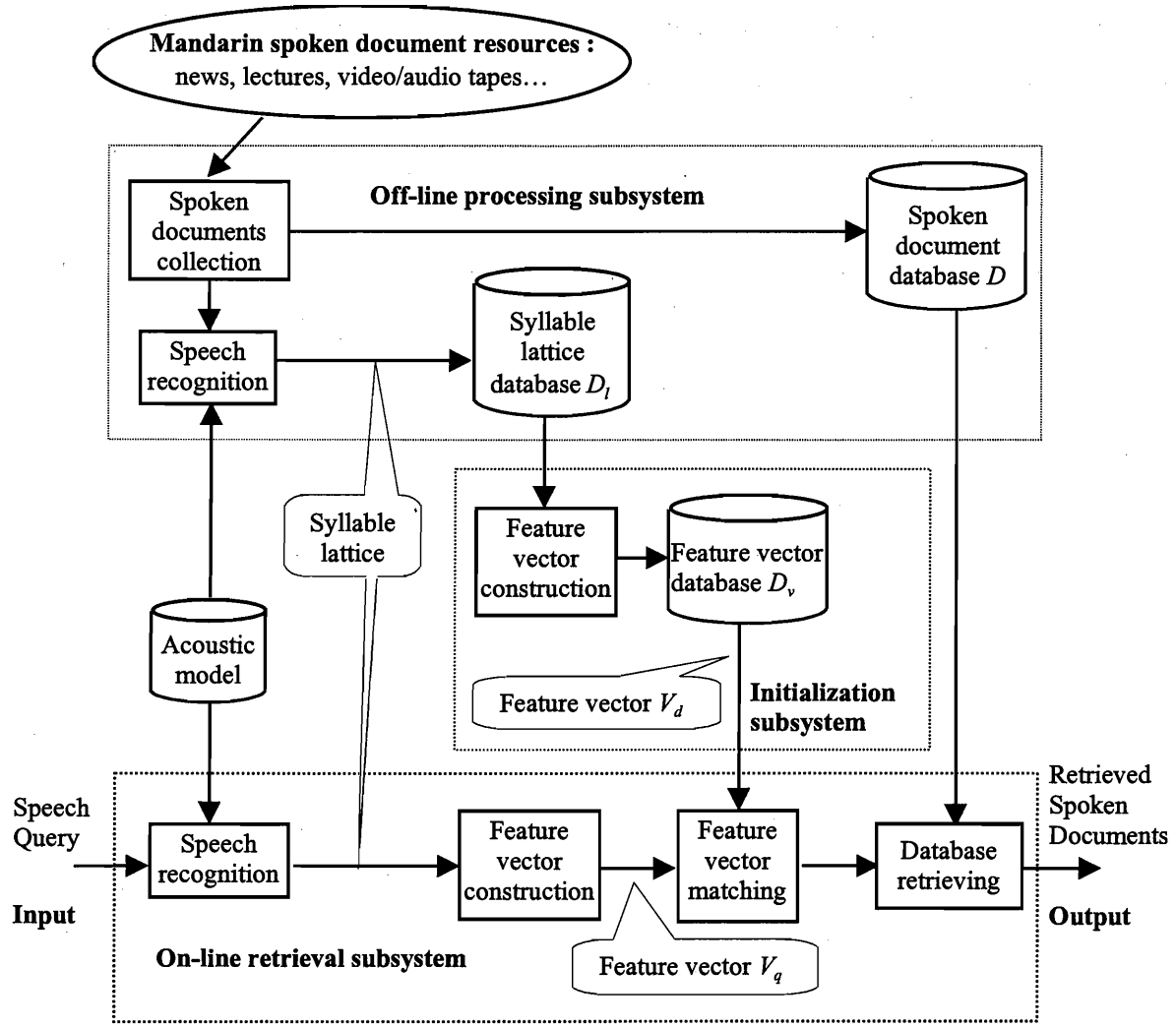


Figure 1: The overall architecture of the previous syllable-based approach for retrieving Mandarin spoken documents using speech queries.

The initialization subsystem is to obtain the feature vectors to be used for retrieval from the syllable lattice database D_l. The feature vector of each document contains the presence information, frequency counts, and acoustic scores of all syllables and adjacent syllable pairs in the syllable lattice. After the feature vectors have been constructed for all syllable lattices in the syllable lattice database D_l, the feature vector database D_v is established, which will be the target database to be physically retrieved. The whole process is also performed off-line in advance.

In the on-line retrieval subsystem, when a speech query is entered, speech recognition will first generate a syllable lattice for the speech query, and then the corresponding feature vector V_q will be constructed based on this syllable lattice via exactly the same processing

procedures as those for spoken documents. Given the feature vector database D_v and the query feature vector V_q, the retrieving module then evaluates the similarity measure between V_q and all feature vectors of the database, and selects a set of documents with the highest similarity measures as the retrieving output.

3.2 Feature Vectors

For each spoken document d in the database D, through searching the syllable lattice, all acoustic scores of single syllables and adjacent syllable pairs in the syllable lattice can be extracted to form the feature vector V_d,

$$V_d = (as(s_1),...,as(s_i),...,as(s_{416}),as(s_1,s_1),...,as(s_i,s_j),...,as(s_{416},s_{416})) \quad (2)$$

where $as(s_i)$ is the acoustic score of the syllable s_i, and $as(s_i,s_j)$ is the acoustic score of the syllable pair (s_i,s_j). The feature vector constructing procedures were performed off-line on all documents in the database D to form a feature vector database D_v, which will be the target database to be physically retrieved. While regarding a query, the same feature vector constructing procedures must be performed on-line to construct the feature vector V_q right after the input query is entered.

3.3 Retrieving Process

Given the feature vector database D_v and a query q, the retrieving problem is actually a searching process to retrieve the document d^* in the target database D_v which is most related to the query. This searching process thus can be formulated as follows:

$$d^* \equiv \arg\max_{d \in D_v} Sim(d,q) \quad (3)$$

where *Sim(d,q)* is a similarity measure between a document d and the query q, and the Cosine measure (Salton, 1983) can be used to estimate the similarity:

$$Sim(d,q) = \cos(V_d,V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|} \quad (4)$$

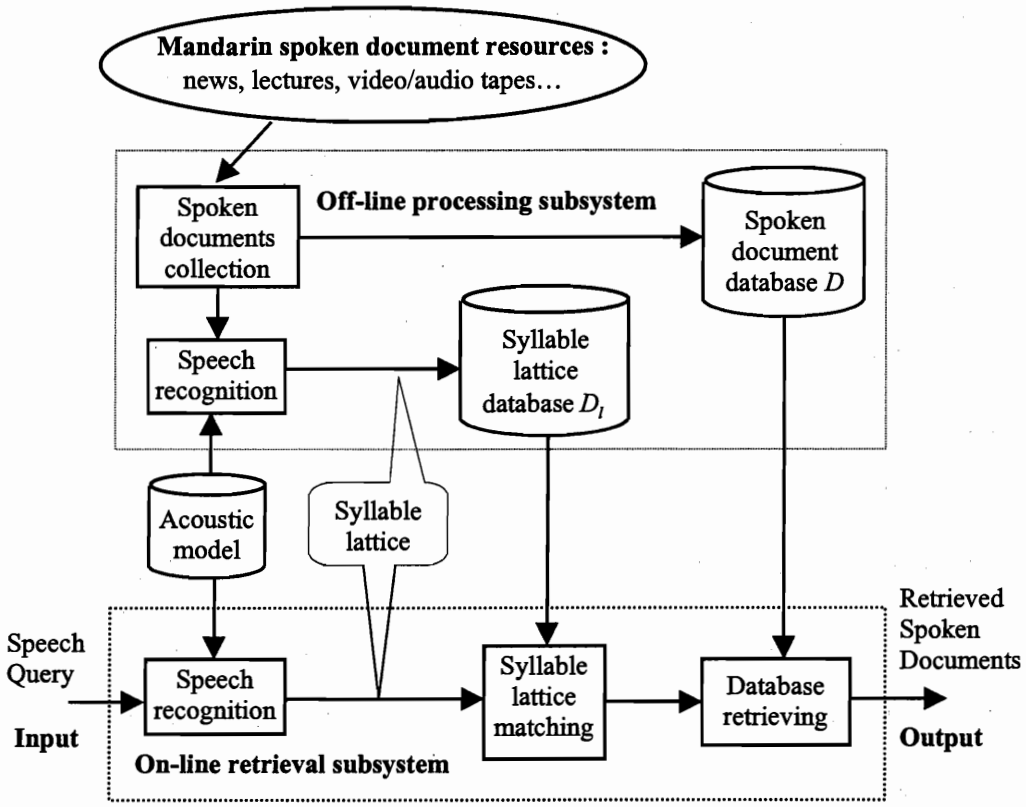


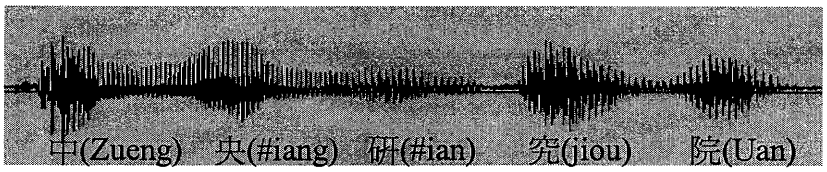
Figure 2: The overall architecture of the new syllable-based approach for retrieving Mandarin spoken documents using speech queries.

4. The New Approach

This section will briefly introduce the overall system architecture and the syllable lattice matching process of the proposed new syllable-based approach.

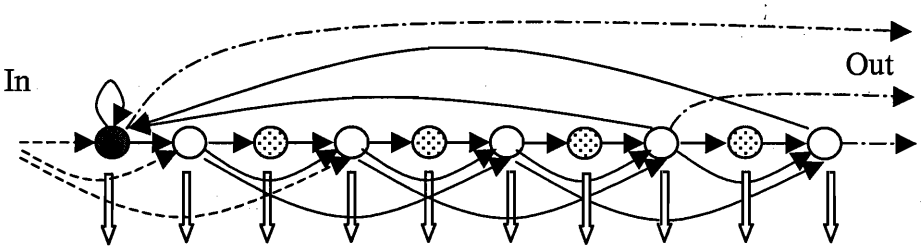
4.1 Overall System Architecture

The overall architecture of the new syllable-based approach for Mandarin spoken document retrieval is shown in Figure 2. The whole system is now divided into two parts. The first part in the upper dotted square of Figure 2 is still the off-line processing subsystem, which is exactly the same as the first part of the previous approach as shown in Figure 1. 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. Note that here the similarity measure is based on directly matching the syllable lattice, the feature vector construction module in the previous approach is therefore no more necessary.



Zueng	jiang	#ian	#iou	Uan
zueng	#iang	#iang	jiou	
Cueng		Uan		

(a) syllable lattice



Observation:	1	375	1	276	1	255	1	247	1	338
	2	379	2	275	2	275	2	248	2	
	3	376	3		3	338	3		3	
	:		:		:		:		:	
	416		416		416		416		416	

(b) DHMM representation of a syllable lattice

Figure 3: An example syllable lattice of a key-phrase speech query "中央研究院(Academia Sinica)", and the corresponding DHMM representation.

4.2 Retrieving Process

Given the syllable lattice database D_l and a query q, the retrieving problem is still a searching process to retrieve the document d^* in the target database D_l which is most related to the query as defined in equation (3). However, here we need a new method to evaluate the similarity measure between a document d and the query q.

As shown in Figure 3, the syllable lattice of a key phrase speech query can be represented as a discrete Hidden Markov Model (DHMM), $\lambda_q = (A, B, \pi)$ (Rabiner and Juang, 1993), 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 of the length (i.e., the syllable number) of the speech query. The first state (the dark one, as shown in Figure 3) 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 in the spoken documents and deletion errors in 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 in the spoken documents and insertion errors in 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 3, e.g. $\pi_i = 1/3, i = 1,2,4$ while $\pi_i = 0, i = 3, \text{or} 4 < i \le N$ 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.

First of all, we can use the Viterbi search algorithm to find the best state sequence. The complete procedure is stated as follows (Rabiner and Juang, 1993):

1.Initialization

$$\delta_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N \tag{5a}$$

$$\psi_1(i) = 0, \qquad 1 \le i \le N \tag{5b}$$

2.Recursion

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(o_t), \qquad 2 \le t \le T, 1 \le j \le N \tag{6a}$$

$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}], \qquad 2 \le t \le T, 1 \le j \le N \tag{6b}$$

3.Termination

$$P^* = \max_{1\le i\le N}[\delta_T(i)], \tag{7a}$$

$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)], \tag{7b}$$

4.State sequence backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1 \tag{8}$$

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:

$$b_j(o_t) = \sum_{k=1}^{K} b_j(o_{tk}) \times as(o_{tk}) \tag{9}$$

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 and estimate the similarity measure between a spoken document *d* and the speech query *q* using the following equation:

$$Sim(d,q) = \sum_{i=1}^{MSN} matched\_score(i) \tag{10}$$

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

$$match\_score(i) = \sum_{t=t_i}^{t_i+D_i-1} b_{q_t}(o_t) \qquad (11)$$

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 retrieving results.

5. Experiments and Discussions

5.1 Speech Database Used in the Experiments

The example speech database to be retrieved in the following experiments consists of 500 Mandarin spoken documents for Chinese news. They were produced by 5 different male speakers. The text materials are news articles published in Taiwan area in 1997. On average, each spoken document contains about 100 characters (i.e., 100 syllables), while the individual length of the articles ranges from 44 to 269 characters. A set of 80 simple key phrase queries produced by 4 different male speakers were used for tests. 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 by searching through the original text materials. 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 2 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.

5.2 Experimental Results

5.2.1 Comparison between the two approaches

The first experiment was tested to make comparison between the previous syllable-based approach and the new syllable-based approach proposed in this paper. The non-interpolated

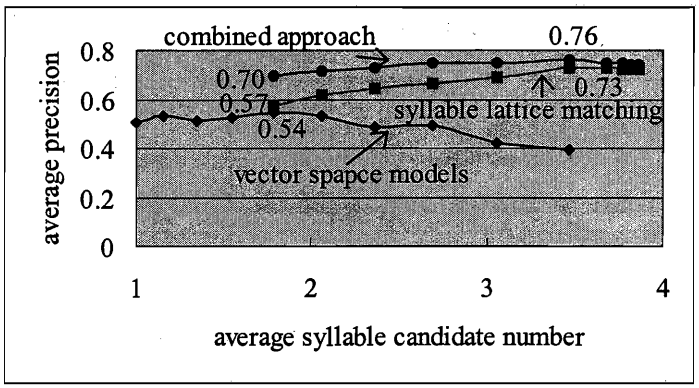


Figure 4: Results for retrieving Mandarin spoken documents using key phrase speech queries based on vector space models, syllable lattice matching, and the combined approach.

average precision rates with respect to the average number of syllable candidates are plotted in Figure 4. For the previous 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 precision rate achieved is 0.54 when the average number of syllable candidates is only 1.79. This result seems counter-intuitive at the first sight. 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 therefore inevitably increase the degree of ambiguity. Although the acoustic recognition scores $as(s_i, s_j)$ in the feature vectors can provide some degree of discrimination against less reliable syllable candidates, but the extra correct syllables included by the increase of the number of syllable candidates very often also have relatively lower acoustic scores. Therefore the information regarding the desired documents carried by these extra correct syllables may be easily swamped by syllable pairs constructed with wrong syllables with relatively higher acoustic recognition scores. These explain why the performance degrades with increased number of syllable candidates. So good retrieving approaches should be able to make use of the increased correct syllables to achieve better results. For the proposed 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 rate achieved is 0.73 when the average number of syllable candidates is 3.47. The new approach produced 0.19 (0.73-0.54) improvements in non-interpolated average precision, while the average number of syllable candidates in this case is almost twice (3.47/1.79=1.94) as the average number of syllable candidates used in the best case of the previous approach. It can also be

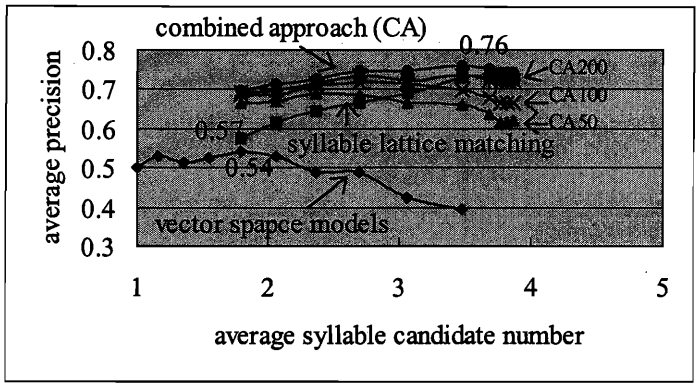


Figure 5: Results for retrieving Mandarin spoken documents using key phrase speech queries based on the two-stage approach (CA200, CA100, and CA50).

found that, for the proposed lattice-matching-based approach, the curve keeps relatively flat as the average number of syllable candidates further increases. The experimental results show that the new approach is better than the previous approach in making use of the syllable lattice, and thus the retrieval performance is significantly improved.

5.2.2 Results for the combined approach

Based on the above descriptions of the methodologies of the two syllable-based approaches and the experimental results, one may wonder how many further improvements can be obtained if we combine the above two approaches together. Since both the vector-space-based and lattice-matching-based approaches are based on exactly the same speech recognition and syllable lattice construction front-end. They can thus be very easily combined together as a combined approach. That is, the similarity measures obtained using equations (4) and (10) can be summed together to give a new similarity measure between a spoken document and a query. The top curve in Figure 4 shows that, in any case, the results for the combined approach are better than the results for either the vector-space-based approach or the lattice-matching-based approach. The best average precision rate achieved is 0.76, while the best average precision rates are 0.73 and 0.54 for the lattice-matching-based approach and the vector-space-based approach respectively.

Furthermore, 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. In fact, in our experiments, the search time for the lattice-matching-based approach was about 10 times of that for the vector-space-based approach. It is thus a good idea to modify the combined approach to a two-stage search strategy for shortening the total search time. In the

first stage, the vector-space-based approach can be 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 these potential documents only. Finally, the potential documents are re-ranked based on the summation of two similarity measures obtained by two approaches and the final results can be obtained. Figure 5 shows the results for this two-stage approach, in which the three curves marked by "CA200", "CA100", and "CA50" represent the cases that the lattice-matching-based approach was applied to 200, 100, and 50 potential documents selected by the vector-space-based approach respectively. Because the retrieval performance of the first-stage vector-space-based approach is relatively poor, more potential documents are therefore necessary to cover the desired documents. This is why, as shown in Figure 5, the performance gets worse and worse with less potential documents applied in the second-stage lattice-matching-based approach. However, a very important result from Figure 5 is that the retrieval performance for the CA50 case (0.69) is still much better than that for using the vector-space-based approach only (0.54). But, in this case, the search time is only about 2 ($1+10\times 50/500$) times of that for using the vector-space-based approach only.

5.3 Discussions

Currently, key phrase queries are widely used in text-based retrieval such as Internet search engines. In fact, key phrase queries are simple, convenient, and efficient. On the other hand, natural language queries inevitably cause more ambiguities and thus degrade the retrieval performance significantly. This is why this paper focuses on retrieval of Mandarin spoken documents using short key phrase speech queries. The experimental results indicate that the proposed approach can significantly improve the retrieval performance as compared to the previous approach. However, the previous approach can be directly applied to spoken document retrieval using natural language speech queries though the retrieval performance is even worse (Bai et al., 1996, 1999), but whether this new approach can be applied to spoken document retrieval using natural language speech queries is yet to be further investigated.

6. Conclusion

In this paper, we propose a new syllable-based approach for retrieving Mandarin spoken documents using short speech queries. This approach that is primarily based-on matching the whole syllable lattice directly can better make use of the syllable lattice obtained by speech recognition as compared to the previous syllable-based approach that using syllable and syllable-pair information extracted from the syllable lattice based on the vector space model.

The experimental results show that the retrieval performance can be significantly improved.

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

Bai, B. R., Chien, L. F., and Lee, L. S. (1996), "Very-large-vocabulary Mandarin voice message file retrieval using speech queries", *Proc. International Conference on Spoken Language Processing*, pp. 1950-1953.

Bai, B. R., Chen, B., and Wang, H. M. (1999), "Syllable-based Chinese text/spoken document retrieval using text/speech queries", *Proc. International Conference on Multimodal Interface*, pp. II46-II51.

CMU Informedia Digital Video Library project http://informedia.cs.cmu.edu/.

Glavitsch, U. and Schäuble, P. (1992), "A system for retrieving speech documents", *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, pp. 168-176.

Harman, D. (1995), "Overview of the Fourth Text Retrieval Conference (TREC-4)", Available at http://trec.nist.gov/pubs/trec4/t4_proceedings.html .

James, D. A. (1995), The application of classical information retrieval techniques to spoken documents, Ph.D. Dissertation, University of Cambridge, UK.

Jones, K. S., Jones, G. J. F., Foote, J. T., and Young, S. J. (1996), "Experiments on spoken document retrieval", *Information Processing & Management*, Vol. 32, No. 4, pp. 399-417.

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

Ng, K. and Zue, V. (1997), "Subword unit representations for spoken document retrieval", *Proc. European Conference on Speech Communication and Technology*, pp. 1607-1610.

Rabiner, L. and Juang, B. H. (1993), *Fundamentals of Speech Recognition* (Prentice-Hall International, Inc.).

Salton, G. (1983), *Introduction to Modern Information Retrieval* (McGraw-Hill, NY).

Wang, H. M. et al. (1997), "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data", *IEEE Trans. Speech and Audio Processing*, Vol. 5, No. 2, pp. 195-200.

Wechsler, M. (1998), Spoken document retrieval based on phoneme recognition, Ph.D. Dissertation, Swiss Federal Institute of Technology (ETH), Zurich.