Content-based Language Models for Spoken Document Retrieval

Hsin-min Wang and Berlin Chen

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

Email: whm@iis.sinica.edu.tw, berlin@iis.sinica.edu.tw

Abstract

Spoken document retrieval (SDR) has been extensively studied in recent years because of its potential use in navigating large multimedia collections in the near future. This paper presents a novel concept of applying content-based language models to spoken document retrieval. In an example task for retrieval of Mandarin Chinese broadcast news data, the content-based language models either trained on automatic transcriptions of spoken documents or adapted from baseline language models using automatic transcriptions of spoken documents were used to create more accurate recognition results and indexing terms from both spoken documents and speech queries. We report on some interesting findings obtained in this research.

Keywords: spoken document retrieval (SDR); content-based language models; speech recognition.

1. Introduction

Massive quantities of spoken audio are becoming available on the web. Therefore, intelligent and efficient information retrieval techniques allowing easy access to spoken documents, such as the radio and television shows, are becoming highly desired and have been extensively studied in recent years [1-6]. There have been several different approaches developed for spoken document retrieval (SDR). Word-based retrieval approaches [1-3] have been very popular and successful, though with the potential problems of either having to know the query words in advance, or requiring a large enough lexicon to cover the growing dynamic contents, such as the diverse broadcast news. Some other researchers [4-6] proposed the concept of subword-based approaches because the subword units could provide a complete phonological coverage for spoken documents and circumvent the OOV problem in audio indexing and retrieval.

In either approach, applying language models to speech recognition can definitely improve recognition accuracy as well as retrieval performance. Lin et al [7] have proved, in their evaluation of retrieving a very large collection of Mandarin Chinese text documents with Mandarin speech queries, that applying the language models trained on the target text database to the recognition of speech queries could significantly improve both recognition accuracy and retrieval performance. However, for spoken document retrieval, the real issue is how to collect text corpora adequate to training the language models. For example, in automatic transcription of broadcast news, many experimental results have shown that using the language models trained from the broadcast news text can achieve higher recognition accuracy compared to using the language models trained from the newswire text [8]. Therefore, there is reason to believe that the language models adapted from baseline language models using the automatic transcriptions of the large collection of spoken

documents may be helpful for both speech recognition and information retrieval. However, for some tasks, such as retrieval of recordings of meetings or voice mails, the adequate corpora may be very difficult to collect or even not available at all. In these cases, only the language models trained on the automatic transcriptions of the large collection of spoken documents are available and they may be helpful as well. Accordingly, this paper presents a concept of applying the content-based language models to spoken document retrieval. The content-based language models can be either trained on automatic transcriptions of spoken documents or adapted from baseline language models using automatic transcriptions of spoken documents. The recognition of spoken documents can be iteratively executed and, thus, more accurate recognition results can be obtained and used to create the indexing terms. Such content-based language models can be used to improve the recognition accuracy of speech queries as well. Hopefully, retrieval performance can be improved correspondingly.

Considering the monosyllabic structure of the Chinese language, a whole class of indexing features for retrieval of Mandarin Chinese broadcast news using syllable-level statistical characteristics has been previously investigated [6]. The syllable-based approach is a special case of subword-based approaches. The feasibility of the proposed approach has been tested on the same task. The rest of this paper is organized as follows. The methodology of the proposed approach to spoken document retrieval is introduced in Section 2. The speech recognition process and the retrieval method are presented in Sections 3 and 4, respectively. Then, the target database to be retrieved is introduced in Section 5, and all the experimental results are discussed in Section 6. Finally, the concluding remarks are made in Section 7.

2. Methodology

The block diagram of the proposed approach to spoken document retrieval is shown in Figure 1. During the phase of database preparation, the speech recognition module first transcribes every spoken document in the database to a word (or subword) string based on the acoustic models. Then, the automatic transcriptions of all the spoken documents are used to train the so-called content-based language models, and the speech recognition module repeats recognition based on the acoustic models and the content-based language models. For a bi-gram model, the bi-gram probabilities are estimated using the following equation

$$p(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}. \quad (1)$$

Where $c(w_{i-1}, w_i)$ and $c(w_{i-1})$ respectively denote the occurrence counts of the word pair (w_{i-1}, w_i) and the word w_{i-1} in the spoken document collection. The details of speech recognition based on the statistical acoustic models and N-gram language models can be found in many books and papers [9]. The language model training and speech recognition processes can iterate several times and higher recognition accuracy can be achieved. If the baseline language models are available, the automatic transcriptions of the spoken documents are used to adapt the baseline language models to the content-based language models. For a bi-gram model, the adapted bi-gram probabilities can be obtained using the following equation

$$p(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) \times \alpha + c_0(w_{i-1}, w_i)}{c(w_{i-1}) \times \alpha + c_0(w_{i-1})}. \quad (2)$$

Where $c_0(w_{i-1}, w_i)$ and $c_0(w_{i-1})$ respectively denote the occurrence counts of the word pair (w_{i-1}, w_i) and the word w_{i-1} in the text corpora on which the baseline language models were

trained, and α is a weighting factor. In this paper, α is simply set to 1. Again, the language model adaptation and speech recognition processes can iterate several times and higher recognition accuracy can be achieved. Finally, the feature vector extraction module constructs the feature vectors from these automatic transcriptions.

When a user enters a speech query into the retrieval system, the speech recognition module first transcribes the speech query to a word (or subword) string based on the acoustic models for speech queries and the content-based language models obtained in the database preparation phase. Then, the feature vector extraction module constructs the feature vector from the word (or subword) string. Finally, the retrieval module evaluates the similarity measures between the feature vector of the speech query and all the feature vectors of the spoken documents and selects a set of documents with the highest similarity measures as the retrieval output.

In this paper, a Mandarin spoken document or a speech query is transcribed to a syllable string and the feature vector consists of overlapping syllable N-grams (uni-gram, bi-gram, and tri-gram) and syllable pairs separated by n syllables ($n=1,2,3$). The details for speech recognition and information retrieval will be given in Sections 3 and 4, respectively.

3. Speech Recognition

3.1 A brief introduction to Mandarin Chinese

In the Chinese language, there are more than 10,000 characters. A Chinese word is composed of one to several characters. The combination of one to several of such characters gives an almost unlimited number of words, which can be found in different versions of dictionaries and texts on different subjects. Two nice features of the language are that all characters are monosyllabic, and that the total number of phonologically allowed syllables is only 1,345. Furthermore, Mandarin

Chinese is a tonal language, in which each syllable is assigned a tone and there are a total of 4 lexical tones plus 1 neutral tone. If the differences among the syllables caused by tones are disregarded, only 416 “base syllables” (i.e., the tone-independent syllable structures) instead of 1,345 different “tonal syllables” (i.e., including the tone features) are required to cover all the pronunciations for Mandarin Chinese [10]. Base syllable recognition is, thus, believed to be the first key problem in the speech recognition as well as in the spoken document retrieval for Mandarin Chinese.

3.2 Signal processing

For speech recognition, typically, a speech signal is divided into a number of overlapping time frames, and a speech parametric vector is computed to represent each time frame. In our speech recognizer, spectral analysis is applied to a 20 msec frame of speech waveform every 10 msec. For each speech frame, 12 mel-frequency cepstral coefficients (MFCC) and the logarithmic energy are extracted, and these 13 coefficients along with their first and second time derivatives are combined together to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to all the training sentences, spoken documents and speech queries.

3.3 Acoustic modeling

Although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of the Chinese language, using it leads to inefficient utilization of the training data in the training phase and high computation requirements in the recognition phase. Thus, the acoustic units chosen here are 112 context-dependent INITIALs and 38 context-independent FINALs based on the monosyllabic nature of Mandarin Chinese and the INITIAL-FINAL structure

of Mandarin base syllables. Here, INITIAL means the initial consonant of the base syllable, and FINAL means the vowel (or diphthong) part but including optional medial or nasal ending.

Each INITIAL is represented by a HMM with 3 states while each FINAL is represented by one with 4 states. The Gaussian mixture number per state ranges from 2 to 16, depending on the amount of training data. Therefore, every syllable unit is represented by a 7-state HMM. The silence model is a 1-state HMM with 32 Gaussian mixtures trained using the non-speech segments.

3.4 Language modeling

The baseline syllable language models are trained on a newswire text corpus consisting of 50M Chinese characters collected from Central News Agency (CNA) in 1999 from January to July. The training materials are word identified and phonetic spelling indicated using a pronunciation lexicon consisting of around 62,000 frequently used Chinese words.

3.5 Syllable recognition

If the language models are not available, the syllable recognizer performs only free syllable decoding without any grammar constraints. On the other hand, in a more complicated syllable recognizer, a two-pass search strategy is used. In the first pass, the Viterbi search is performed based on the acoustic models and the syllable bi-gram language models, and the score at every time index is stored. In the second pass, a backward time-asynchronous A* tree search [11] generates the best syllable string based on the heuristic scores obtained from the first pass search and the syllable bi-gram language models.

4. Information Retrieval

4.1 Vector space models

Vector space models widely used in many text information retrieval systems are used here, in which each document or query is represented by a feature vector:

$$V = (w_1 \times f(t_1) \times idf(t_1), \ldots, w_k \times f(t_K) \times idf(t_K)) . \quad (3)$$

Where w_i, $f(t_i)$ and $idf(t_i)$ are respectively the weight, frequency, and inverse document frequency (IDF) of the indexing term t_i, while *K* is the total number of distinct indexing terms. Here, the term frequency is defined as follows:

$$f(t_i) = 1 + \ln\left(\sum_{j=1}^{j=n_{t_i}} cm_{t_i}(j)\right) . \quad (4)$$

Where $cm_{t_i}(j)$, ranging from 0 to 1, is the normalized acoustic confidence measure of the *j*-th occurrence of the indexing term t_i in the document or query, and n_{t_i} is the total occurrences of the indexing term t_i in the document or query.

The weight w_i is different for different classes of indexing terms as explained below. The Cosine measure widely used in text information retrieval is used to estimate the similarity between a document and a query.

4.2 Syllable-level indexing terms

In this paper, the syllable-level indexing terms compose of overlapping syllable segments with length *N* (*S(N), N=1~3*) and syllable pairs separated by *n* syllables (*P(n), n=1~3*). Here, the syllable segment with length *N* corresponds to the overlapping syllable N-gram. Considering a syllable

string of 10 syllables S_1 S_2 S_3.... $S_{10,}$ examples of the former are listed on the upper half of Table 1, while examples of the latter on the lower half of Table 1. The combination of these indexing terms has been shown very effective for Mandarin Chinese SDR [6]. For example, the overlapping syllable segments with length *N* can capture the information of polysyllabic words or phrases while the syllable pairs separated by *n* syllables can tackle the problems arising from the flexible wording structure, abbreviation, and speech recognition errors. In the following experiments, the weight in Equation (3) is 0.1 for *S(1)* and 1.0 for all the rest classes of indexing terms.

5. Data collection

In this paper, the radio news was recorded using a wizard FM radio connected to a PC, and digitized at a sampling rate of 16kHz with 16bit resolution. The data were collected from several radio stations, all located at Taipei, from December 1998 to July 1999. All the recordings were manually segmented into stories and transcribed to texts.

The target database to be retrieved in the following experiments consists of 757 recordings (about 10.5 hours of speech materials). They were collected from Broadcasting Corporation of China (BCC). Each recording is a short news abstract (about 50 seconds of speech materials) produced by a news announcer. Hence, each recording may contain several news items. Some recordings in the database contain background music. 40 simple queries were manually created to support the retrieval experiments. Each query has on average 23.3 relevant documents among the 757 documents in the database, with the exact number ranging from 1 to 75. Two (one male and one female) speakers were asked to record the 40 queries, respectively. At the first glance, this SDR task seems easy since the retrieval target contains only anchors' speech. However, this SDR task is, in fact, very difficult because almost all the query terms appear only once in each of their relevant

documents and the queries are often very short. On average, each query contains roughly only 4 characters (or syllables).

A different broadcast news speech database consisting of 453 stories (about 4.0 hours of speech materials) collected from Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT) is used for training the speaker-independent HMMs for the recognition of broadcast news speech. Another read speech database including 5.3 hours of speech materials for phonetically balanced sentences and isolated words produced by roughly 80 male and 40 female speakers is used for training the speaker-independent HMMs for the recognition of speech queries.

6. Experiments

We have evaluated the proposed approach in two different ways. In the first way, we iteratively perform syllable recognition on all the spoken documents. On the other hand, in the second way, we perform syllable recognition on all the spoken documents just once and every spoken document is represented as a syllable lattice first. Then, we iteratively perform the re-scoring process on every syllable lattice based on the acoustic recognition scores and the content-based language models to generate a new best syllable string. The experimental results show that the proposed approach of applying content-based language models to speech recognition in either way not only improves the recognition accuracies of both the spoken documents and the speech queries but also improves the retrieval performance.

6.1 Experimental results based on the standard approach

As described in Section 2, in the database preparation phase, we can first compute the content-based language models using the automatic transcriptions of all the spoken documents using Equations (1)

or (2) and, then, re-perform the syllable recognition process on all the spoken documents. The above process can be performed iteratively and, hopefully, both recognition accuracy and retrieval performance can be improved.

6.1.1 Syllable recognition results

Table 2 summarizes the syllable recognition accuracies obtained in this manner. In the first experiment, we assume that the baseline language models are not available. In the second row of Table 2, the test using free syllable decoding gives the recognition accuracy of spoken documents only 56.11%. With the content-based syllable bi-gram language models applied, the recognition accuracy can be improved to 60.77%, as shown in the third row. That is, with the content-based language models applied, the recognition accuracy is improved by about 8%. These results indicate that the concept of applying content-based language models to the recognition of spoken documents is useful in improving recognition accuracy. However, the system has to perform speech recognition twice, which is a very time consuming task. We have also tested on the speech queries. The test using free syllable decoding gives the recognition accuracy only 55.56%. With the content-based language models applied to the recognition of speech queries, the recognition accuracy can be improved to 70.68%. These results show that the content-based language models are even more useful to the recognition of speech queries. The recognition accuracy is improved by about 27%.

If we apply the baseline language models to the initial recognition of spoken documents and speech queries, the recognition accuracies are 64.68% and 72.84%, respectively, as shown in the fourth row of Table 2. With the content-based language models adapted from the baseline language models using automatic transcriptions of spoken documents applied to speech recognition, the recognition accuracies are 64.52% and 72.84%, respectively. The accuracies almost remain

unchanged, probably because the weighting factor α used in Equation (2) was not carefully chosen. α is set to 1.0 for simplicity at the moment. However, the text corpus for training the baseline language models contains about 50M characters (syllables) while the spoken document collection contains less than 200K characters (syllables). The occurrence counts of syllables and syllable pairs contributed by the relatively small spoken document collection can hardly influence the total occurrence counts in Equation (2). Hence, designing a sophisticated formula or procedure to manipulate the value of α, according to the size of the text corpus used for training the baseline language models, the size of the spoken document collection, and speech recognition accuracy or the acoustic recognition scores of words (We used syllables in this paper.), is worthy of further study. A sophisticated equation for language model adaptation may help as well. Furthermore, in typical speech recognition systems, speaker adaptation techniques are widely used to customise the HMMs to the characteristics of a particular speaker using a small amount of enrollment or adaptation data. Hence, the other possible ways include applying the speaker adaptation techniques to language model adaptation and integrating the language model adaptation techniques with the speaker adaptation techniques.

6.1.2 Spoken document retrieval

Retrieval performance in terms of non-interpolated average precision [12] is shown in Table 3, where SD/SQ and SD/TQ represent the results of spoken document retrieval obtained using speech queries and text queries, respectively. That is, the erroneous syllable strings of documents and queries obtained from speech recognition are denoted as SD (Spoken Documents) and SQ (Speech Queries), respectively, while the query transcripts, which are exactly correct, are denoted as TQ (Text Queries). Again, we first assume the baseline language models are not available. The non-

interpolated average precisions for SD/SQ and SD/TQ based on the transcriptions obtained by free syllable decoding are 0.2524 and 0.4366, respectively, as shown in the second row of Table 3. With the content-based syllable bi-gram language models trained on automatic transcriptions of spoken documents applied to speech recognition, they are improved to 0.3617 and 0.4732, respectively, as shown in the third row.

If we apply the baseline language models to the recognition of spoken documents and speech queries, the non-interpolated average precisions for SD/SQ and SD/TQ are 0.4264 and 0.5541, respectively, as shown in the fourth row of Table 3. With the content-based language models adapted from the baseline language models using automatic transcriptions of spoken documents applied to speech recognition, the average precisions are 0.4264 and 0.5562, respectively. Again, like recognition accuracies, the retrieval performance almost remained unchanged when the content-based language models were applied to speech recognition. The language model adaptation case is worthy of further study as mentioned above.

6.2 Experimental results based on the fast approach

Because iteratively applying the syllable recognition process to all the spoken documents in the large collection is a very time consuming task, we have designed an alternative way to save the recognition time needed. First of all, all the spoken documents are transcribed to syllable lattices. Then, in each iteration, the content-based syllable bi-gram language models are used to select a new best syllable string from the syllable lattice of a spoken document. The content-based language models can be either adapted from the baseline syllable bi-gram language models using the best syllable strings of all the spoken documents obtained in the previous iteration or directly trained on the best syllable strings obtained in the previous iteration. The above iterative process for finding a

new best syllable string from a syllable lattice based on the acoustic recognition scores and the content-based syllable bi-gram language models is very fast because only a relatively simple re-scoring process instead of a complicated speech recognition process is executed. Such a new syllable string obtained after each iteration is not an optimal result based on the acoustic models and the content-based language models because it is obtained under the constraint of the segmentation obtained from free syllable decoding and the syllable candidates contained in the syllable lattice.

The syllable lattice generation process is described as follows. First of all, the syllable recognizer, which performs free syllable decoding without any grammar constraints or performs syllable recognition based on both acoustic models and baseline language models, is used to transcribe all the spoken documents. Then, based on the state likelihood scores calculated in the first search and the syllable boundaries of the best syllable string, the syllable recognizer further performs the Viterbi search on each utterance segment which may include a syllable and outputs several most likely syllable candidates with their corresponding acoustic recognition scores. Finally, a syllable lattice is constructed. In this study, each syllable segment contains 25 syllable candidates.

6.2.1 Syllable recognition results

Table 4 summarizes the syllable recognition accuracies obtained in this manner. We have first evaluated on training the content-based language models from automatic transcriptions of spoken documents. That is, we assume that the baseline language models are not available. As shown in the second row, the test using free syllable decoding gives the recognition accuracies of spoken documents and speech queries only 56.11% and 55.56%, respectively. With the syllable bi-gram language models trained on the automatic transcriptions of the spoken documents obtained using free syllable decoding applied to re-scoring, the recognition accuracies are improved to 57.53% and

56.48%, respectively, as shown in the third row. With the syllable bi-gram language models trained on the new transcriptions applied to re-scoring, the recognition accuracies are improved to 58.81% and 61.11%, respectively, as shown in the fourth row. After an additional iteration, the recognition accuracies are further improved to 58.91% and 62.04%, respectively, as shown in the fifth row.

We have also evaluated on adapting the baseline language models using automatic transcriptions of spoken documents. With the baseline language models applied to the initial syllable recognition, the recognition accuracies of spoken documents and speech queries are 64.68% and 72.84%, respectively, as shown in the sixth row of Table 4. With the content-based syllable bi-gram language models adapted from the baseline language models using automatic transcriptions of spoken documents applied to re-scoring, the recognition accuracies are 64.51% and 72.84%, respectively, as shown in the seventh row. After one more iteration, the recognition accuracies are 64.49% and 73.15%, respectively, as shown in the last row. The recognition accuracies almost remain unchanged probably due to the same reasons that we have discussed in Section 6.1.1.

The detailed recognition results after the first 10 iterations are depicted in Figure 2. In the case that the baseline language models are not used, the recognition accuracies are improved at first, but, subsequently, the improvements are not obvious or sometimes the recognition accuracy even slightly degrades. In the case with the baseline language models applied, the curves are almost flat.

6.2.2 Spoken document retrieval

Retrieval performance in terms of non-interpolated average precision with respect to the number of iterations is depicted in Figure 3. In general, Figure 3 reveals very similar trends to Figure 2, which displayed the speech recognition accuracies, because the retrieval performance depends strongly on the recognition accuracies of both spoken documents and speech queries. In the case without the

baseline language models applied, the performance improvements for the SD/TQ case are less significant than that for the SD/SQ case. This is because the retrieval performance for the SD/SQ case depends not only on the recognition accuracy of spoken documents but also on the recognition accuracy of speech queries. According to Table 4 and Figure 2, it's obvious that the improvements in the query recognition accuracy are more significant than the improvements in the document recognition accuracy in this task. In the case with the baseline language models applied, the curves for both SD/TQ and SD/SQ are almost flat, just like the curves for the recognition accuracies depicted in Figure 2(b). The best non-interpolated average precisions for SD/SQ and SD/TQ based on the transcriptions obtained by applying the content-based syllable bi-gram language models trained on automatic transcriptions of spoken documents to speech recognition are 0.3340 and 0.4474, respectively, while they are 0.2524 and 0.4366 based on automatic transcriptions obtained by free syllable decoding. The non-interpolated average precisions for SD/SQ and SD/TQ are 0.4264 and 0.5541 based on transcriptions obtained by applying the baseline language models to speech recognition. The best non-interpolated average precisions for SD/SQ and SD/TQ based on transcriptions obtained by applying the content-based syllable bi-gram language models adapted from the baseline language models using transcriptions of spoken documents to speech recognition are 0.4323 and 0.5518, respectively. The retrieval performances after the first several iterations are listed in Table 5.

7. Conclusions

In this paper, we have proposed a novel approach, which applied content-based language models to the recognition of spoken documents and speech queries, to spoken document retrieval. The content-based language models can be either trained on automatic transcriptions of spoken

documents or adapted from baseline language models using automatic transcriptions of spoken documents. We have tested the proposed approach on an example task for retrieval of Mandarin Chinese broadcast news. The experimental results have shown the potential in this direction. Our key findings include:

(i) In the case that the baseline language models are not available, the content-based language models trained on automatic transcriptions of spoken documents are useful to the recognition of both spoken documents and speech queries as well as to the following spoken document retrieval.

(ii) In the case that the baseline language models are available, the situations are more complicated and it is worthy of further study.

(iii) The standard procedure, in which the content-based language models are iteratively applied to the recognition of spoken documents, is very time consuming. On the other hand, the fast procedure needs very little additional time in the database preparation phase but results in reasonable improvements in both recognition accuracy and retrieval performance.

Acknowledgments

The authors wish to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 89-2213-E-001-049.

References

1 Jones, K. S., Jones, G. J. F., Foote, J. T. and Young, S. J. Experiments on spoken document retrieval. Information Processing & Management, 1996, 32(4), pp. 399-417.

2 Wactlar, H., Kanade, T., Smith, M. and Stevens, S. Intelligent access to digital video: the Informedia project. IEEE Computer, 1996, 29(5), pp. 46-52.

3 Voorhees, E. and Harman, D. Overview of the eighth text retrieval conference (TREC-8). In Proceedings of the Eighth Text REtrieval Conference, 1999.

4 Ng, K. and Zue, V. Phonetic recognition for spoken document retrieval. In Proceedings of the 1998 International Conference on Spoken Language Processing, 1998.

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

6 Chen, B., Wang, H. M. and Lee, L. S. Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics. In Proceedings of the 2000 International Conference on Acoustics Speech and Signal Processing, 2000.

7 Lin, S. C., Chien, L. F., Chen, K. J. and Lee, L. S. A syllable-based very-large-vocabulary voice retrieval system for Chinese database with textual attributes. In Proceedings of the 1995 European Conference on Speech Communication and Technology, 1995.

8 Matsuoka, T., Taguchi, Y., Ohtsuki, K., Furui, S. and Shirai, K. Towards automatic transcription of Japanese broadcast news. In Proceedings of the 1997 European Conference on Speech Communication and Technology, 1997.

9 Rabiner, L. and Juang, B. H. Fundamentals of speech recognition. Prentice-Hall International Inc. 1993.

10 Lee, L. S., Voice Dictation of Mandarin Chinese. IEEE Signal Processing Magazine, 14(4), pp. 63-101, 1997.

11 Chen, B., Wang, H. M., Chien, L. F. and Lee, L. S. A*-admissible key-phrase spotting with sub-syllable level utterance verification. In Proceedings of the 1998 International Conference on Spoken Language Processing, 1998.

12 Harman, D. Overview of the Second Text REtrieval Conference (TREC-2). In Proceedings of *TREC-2*, pp. 1-20. Available at http://trec.nist.gov/pubs/trec2/papers/txt/01.txt.

Syllable Segments with Length *N*	Examples
S(N), N=1	*(S_1) (S_2) …(S_{10})*
S(N), N=2	*(S_1 S_2) (S_2 S_3)…(S_9 S_{10})*
S(N), N=3	*(S_1 S_2 S_3) (S_2 S_3 S_4)…(S_8 S_9 S_{10})*
Syllable Pairs Separated by *n* Syllables	Examples
P(n), n=1	*(S_1 S_3) (S_2 S_4) …(S_8 S_{10})*
P(n), n=2	*(S_1 S_4) (S_2 S_5) …(S_7 S_{10})*
P(n), n=3	*(S_1 S_5) (S_2 S_6) …(S_6 S_{10})*

Table 1: The indexing terms extracted from an example syllable string S_1 S_2 S_3…. S_{10}.

	Documents	Queries
Free Syllable Decoding (Without LM)	56.11	55.56
Applying LM Trained on Automatic Transcriptions of BN (Iter 1)	60.77	70.68
Applying LM Trained on Newswire (Baseline LM)	64.68	72.84
Applying LM Adapted by Automatic Transcriptions of BN (Iter 1)	64.52	72.84

Table 2: Syllable recognition accuracies (%) for spoken documents and speech queries, using the standard approach.

	Average Precision (SD/SQ)	Average Precision (SD/TQ)
Free Syllable Decoding (Without LM)	0.2524	0.4366
Applying LM Trained on Automatic Transcriptions of BN (Iter 1)	0.3617	0.4732
Applying LM Trained on Newswire (Baseline LM)	0.4264	0.5541
Applying LM Adapted by Automatic Transcriptions of BN (Iter 1)	0.4264	0.5562

Table 3: Retrieval performance in non-interpolated average precision for spoken document retrieval using speech queries (SD/SQ) and text queries (SD/TQ), using the standard approach.

	Documents	Queries
Free Syllable Decoding (Without LM)	56.11	55.56
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 1)	57.53	56.48
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 2)	58.81	61.11
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 3)	58.91	62.04
Re-scoring with LM Trained on Newswire (Baseline LM)	64.68	72.84
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 1)	64.51	72.84
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 2)	64.51	73.15

Table 4: Syllable recognition accuracies (%) for spoken documents and speech queries, using the fast approach.

	Average Precision (SD/SQ)	Average Precision (SD/TQ)
Free Syllable Decoding (Without LM)	0.2524	0.4366
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 1)	0.3050	0.4443
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 2)	0.3220	0.4458
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 3)	0.3277	0.4465
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 4)	0.3319	0.4474
Re-scoring with LM Trained on Automatic Transcriptions of BN (Iter 5)	0.3340	0.4470
Re-scoring with LM Trained on Newswire (Baseline LM)	0.4264	0.5541
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 1)	0.4272	0.5507
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 2)	0.4293	0.5505
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 3)	0.4281	0.5502

Table 5: Retrieval performance in non-interpolated average precision for spoken document retrieval using speech queries (SD/SQ) and text queries (SD/TQ), using the fast approach.

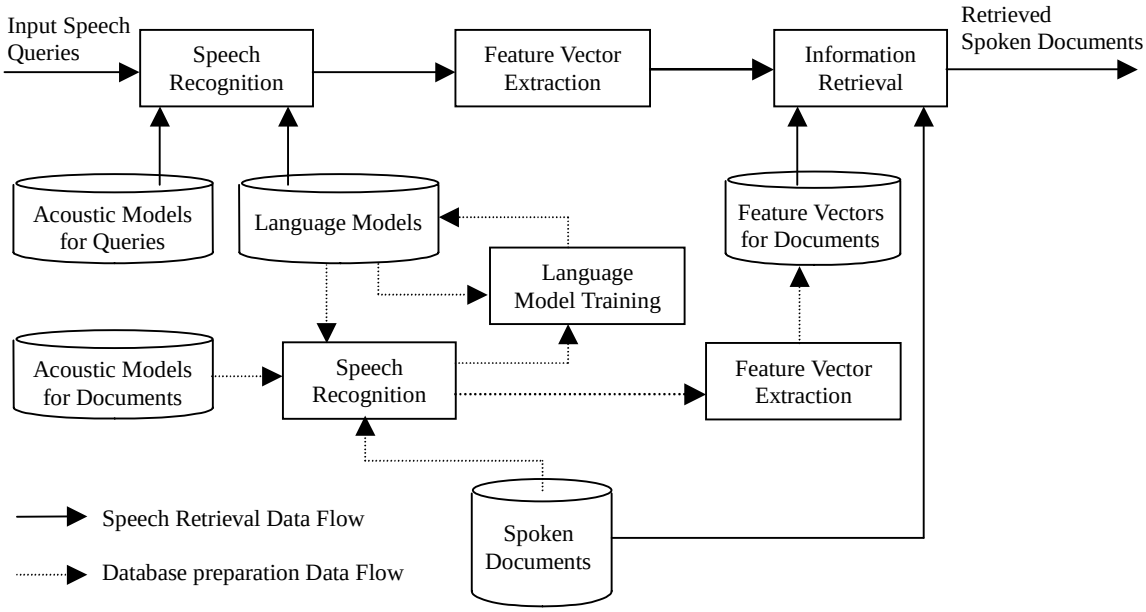
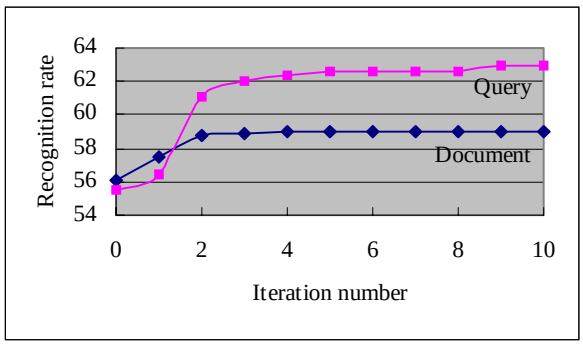
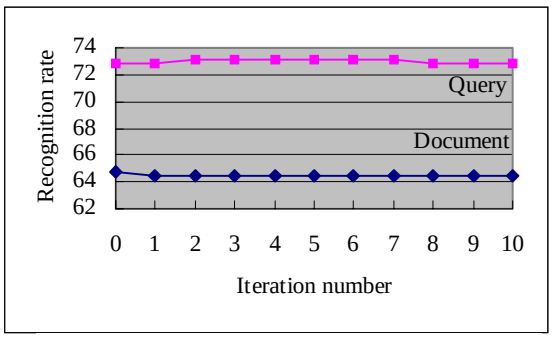


Figure 1: Block diagram of the proposed approach to retrieving spoken documents using speech queries.



(a) Language models were trained on automatic transcriptions



(b) Language models were adapted from baseline language models using automatic transcriptions

Figure 2: Syllable recognition accuracies (%) for spoken documents and speech queries, using the fast approach.

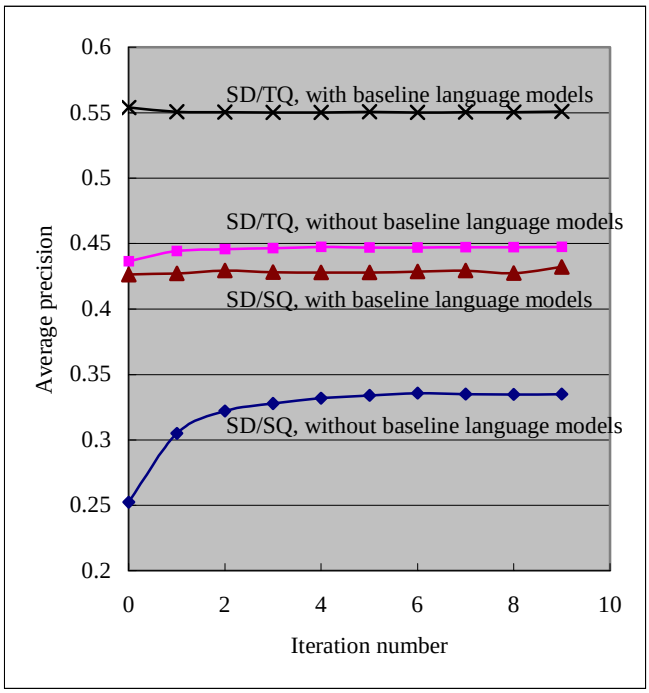


Figure 3: Retrieval performance in non-interpolated average precision for spoken document retrieval using speech queries (SD/SQ) and text queries (SD/TQ), using the fast approach.