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 the content-based language models to spoken document retrieval. In an example task for retrieval of Mandarin broadcast news, the content-based language models either trained with the automatic transcriptions of the spoken documents or adapted from the baseline language models using the automatic transcriptions of the spoken documents were used to create the more accurate recognition results and indexing terms from both the spoken documents and the 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 radio and television programs, 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.

No matter which approach is adopted, applying the language models to speech recognition can definitely improve the recognition accuracy as well as the retrieval performance. Lin et al [7] proved that applying the language models trained on the target text database to recognition of the speech queries could improve the performance of retrieving a very large collection of Chinese text documents with speech queries. However, for spoken document retrieval, the real issue is how to collect adequate corpora for training the language models. For example, in automatic transcription of broadcast news, many experimental results have shown that using the language models trained with the manuscripts of the broadcast news can achieve the higher recognition accuracy compared to using the language models merely trained on the newswire text corpora [8]. Therefore, there is reason to believe that the language models adapted from the baseline language models using the automatic transcriptions of the large collection of spoken documents may be helpful for both recognition and retrieval. However, for some tasks, such as retrieval of meeting notes, the adequate corpora may be very difficult to collect or even not available at all. In these cases, only the language models trained with 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 model can be either trained with the automatic transcriptions of the spoken documents or adapted from the baseline language models using the automatic transcriptions of the spoken documents. The recognition of the spoken documents can be iteratively executed and, thus, the 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 the speech queries as well. As a result, the retrieval performance can be improved correspondingly.

A syllable-based framework for retrieving Mandarin radio broadcast news using speech queries has been investigated in [6]. The syllable-based approach is a special case of the subword-based approaches. This is based on various considerations on the structural nature of the Chinese language. 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, all the experimental results are discussed in Section 5. Finally, the concluding remarks are made in Section 6.

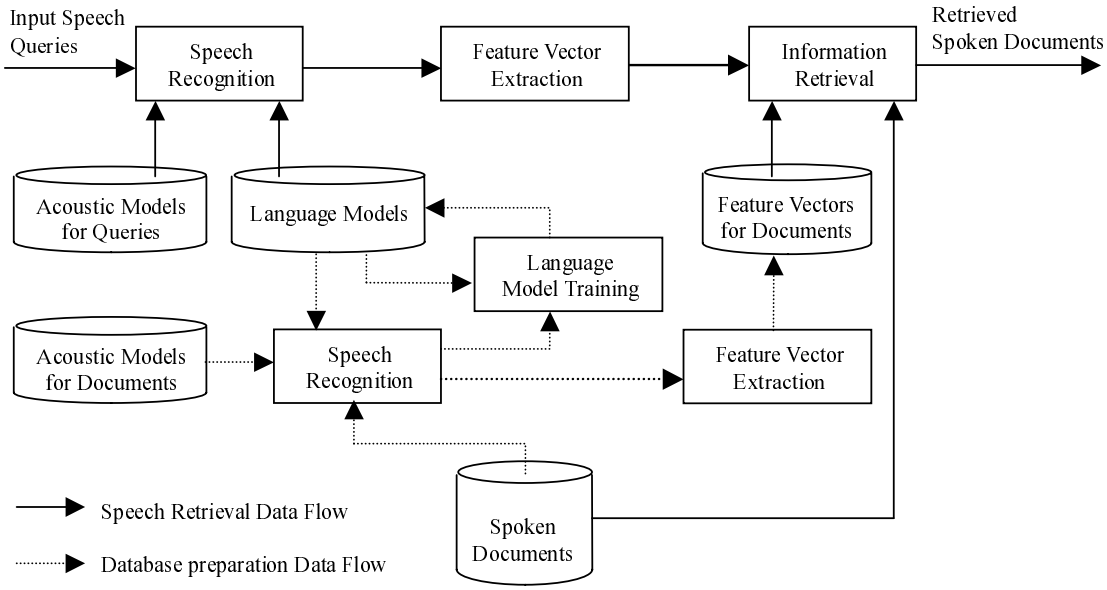


Figure 1: Block diagram of the proposed approach to retrieving spoken documents using speech queries.

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 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

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

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 process can iterate several times and the 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 and the content-based language models are obtained accordingly. For a bi-gram model, the adapted bi-gram probabilities can be obtained using the following equation

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

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 with which the baseline language models were trained, and α is a weighting factor. In this paper, α is simply set to 1. The language model adaptation and speech recognition process can iterate several times and the 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. 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 the overlapping syllable N-grams (uni-gram, bi-gram, and tri-gram) and the 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 Acoustic Processing

Each frame of the speech data is represented by a 39 dimensional feature vector, which consists of 12 mel-frequency cepstral coefficients (MFCCs) and the logarithmic energy, and their first and second time

derivatives. Utterance-based cepstral mean subtraction (CMS) is applied to all the training sentences, spoken documents and speech queries. The acoustic units chosen for syllable recognition are 112 context-dependent INITIALs and 38 context-independent FINALs based on the monosyllabic nature of Mandarin Chinese and the phonetic structure of Mandarin syllables [6]. Here, INITIAL means the initial consonant of the 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. The silence model is a 1-state HMM with 32 Gaussian mixtures trained using the non-speech segments.

3.2 Language Modeling

The baseline syllable language models are trained on a newswire text corpus consisting of 50 million 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 lexicon consisting of around 62,000 frequently used Chinese words.

3.3 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 [10], 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 model, and the score at every time index is stored. In the second pass, a backward time-asynchronous A* tree search generates the best syllable string based on the heuristic scores obtained from the first pass search and the syllable tri-gram language model.

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:

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

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. 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

Syllable Segments	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 Pair 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}.

In this paper, the syllable-level indexing terms compose of the overlapping syllable segments with length N *(S(N), N=1~3)* and the 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 to be very effective for Mandarin 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 Experiments

5.1 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.

The database to be retrieved consists of 757 recordings (about 10.2 hours of speech materials), and was 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, and contains several news items. Some recordings in the database contain background music. A set of 40 simple queries and the corresponding relevant news recordings 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 pronounce the 40 queries, respectively. At the first glance, this SDR task seems easy since the retrieval target contains only the 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

	Documents	Queries
Free Syllable Decoding (Without LM)	56.11	55.56
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 1)	57.53	56.48
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 2)	58.81	61.11
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 3)	58.91	62.04
Re-scoring with LM Trained by Newswire (Baseline LM)	58.44	59.26
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 1)	59.76	63.27
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 2)	59.82	63.27

Table 2: The syllable recognition results for the spoken documents and the speech queries.

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 automatic recognition of the 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 automatic recognition of the speech queries.

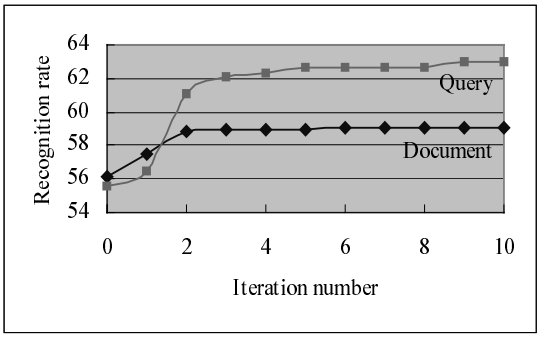
5.2 Syllable Recognition

Because iteratively applying the complicated syllable recognition process, as described previously in Section 3.3, to all the spoken documents in the large database is a very time consuming task, we have chosen an alternative procedure for saving the simulation time in the initial evaluation. First of all, all the spoken documents and speech queries are transcribed to syllable lattices. Then, in each iteration, the content-based syllable bi-gram language model, either adapted from the baseline syllable bi-gram language model 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, is used to select a new best syllable string from the syllable lattice of a spoken document or a speech query. 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 model can be very fast because only a simple re-scoring process instead of a complicated speech recognition process is executed. Obviously, such a new syllable string obtained after each iteration is not an optimal result based on the acoustic models and the content-based language model because it is obtained under the constraint of the segmentation obtained from free syllable decoding and the syllable candidates contained in the syllable lattice. However, such a simplified experimental setup can still show significant improvements in both recognition and retrieval as will be described later on.

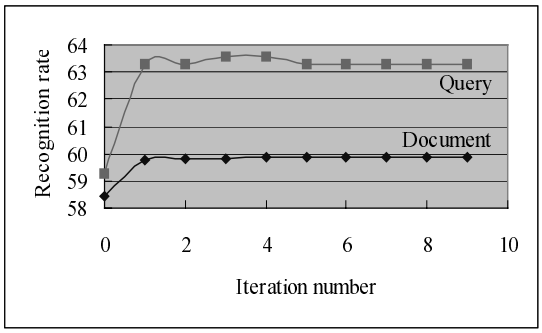
The syllable lattice generation process is described as follows. First of all, the baseline syllable recognizer, which performs only free syllable decoding without any grammar constraints, is used to transcribe all the spoken documents and speech queries. Then, based on the state likelihood scores calculated in the 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 possible syllable candidates with their corresponding acoustic recognition scores. Finally, a syllable lattice is constructed. In this study, each syllable segment contains 25 syllable candidates.

5.2.1 Syllable Recognition Results

Some initial speech recognition experiments have been conducted with the results shown in Table 2. In the second row, the test using free syllable decoding gives the recognition accuracies of the spoken documents and the speech queries only 56.11% and 55.56%, respectively. We have first evaluated on training the content-based language models from the automatic transcriptions of the spoken documents. With the syllable bi-gram language model trained on the transcriptions 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 model 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 the transcriptions of the spoken documents. With the baseline syllable bi-gram language model trained by a newswire text corpus applied to re-scoring, the recognition accuracies can be improved from 56.11% and 55.56% to 58.44% and 59.26%, respectively, as shown in the sixth row of Table 2. With the content-based syllable bi-gram language model adapted from the baseline language model using the new transcriptions applied to re-scoring, the recognition accuracies are further improved to 59.76% and 63.27%, respectively, as shown in the seventh row. However, after one more iteration, the improvements are not obvious, as shown in the last row.



(a) Language models are trained with automatic transcriptions



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

Figure 2: The syllable recognition results for the spoken documents and the speech queries.

	Average Precision (SD/SQ)	Average Precision (SD/TQ)
Free Syllable Decoding (Without LM)	25.24	43.66
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 1)	30.50	44.43
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 2)	32.20	44.58
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 3)	32.77	44.65
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 4)	33.19	44.74
Re-scoring with LM Trained by Automatic Transcriptions of BN (Iter 5)	33.40	44.70
Re-scoring with LM Trained by Newswire (Baseline LM)	31.00	45.50
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 1)	33.68	46.16
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 2)	35.32	46.59
Re-scoring with LM Adapted by Automatic Transcriptions of BN (Iter 3)	35.47	46.49

Table 3: The performance for spoken document retrieval using speech queries (SD/SQ) and text queries (SD/TQ), with/without the baseline language models.

The detailed recognition results are shown in Figure 2. The recognition accuracies are significantly improved at first. Subsequently, the improvements are not obvious or sometimes the recognition accuracy even slightly degrades. One possible reason why the recognition accuracy reaches a saturation point soon is because, in this initial study, the recognizer only performs re-scoring under the constraint of the syllable lattice. There is reason to believe that the higher accuracy can be achieved if the complicated speech recognizer is iteratively applied to recognition of the spoken documents. As a result, a more accurate content-based language model can be obtained based on the more accurate transcriptions of the spoken documents and, thus, the recognition accuracy of the speech queries can be further improved as well. However, the preliminary results have shown the feasibility of using the automatic transcriptions of the spoken documents to adapt or train the content-based language models and, based on which, the recognition accuracies of both the spoken documents and the speech queries can be improved.

5.3 Spoken Document Retrieval

The retrieval performance in terms of non-interpolated average precision [11] with respect to the number of iterations is shown in Figure 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 exactly correct query transcriptions are denoted as TQ (Text Queries). In general, Figure 3 reveals very similar trends to Figure 2 that displayed the speech recognition results; i.e., the improvements in the retrieval performance are significant at first, but not obvious subsequently. This is because the retrieval performance inevitably depends on the recognition accuracies of both the spoken documents and the speech queries. Furthermore, it can be found from Figure 3 that the 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 not only depends on the recognition accuracy of the spoken documents but also highly depends on the recognition accuracy of the speech queries. According to Table 2 and Figure 2, it's obvious that the improvements in the recognition accuracy for the speech queries are more significant than that for the spoken documents in this task. The retrieval performance after the first several iterations is listed in Table 3 in more detail.

The best non-interpolated average precision for SD/SQ and SD/TQ based on the transcriptions obtained by applying the content-based syllable bi-gram language model trained with the transcriptions of the spoken documents to speech recognition is 33.40% and 44.74%, respectively, while it is 25.24% and 43.66% based on the transcriptions obtained by free syllable decoding. Furthermore, the best non-interpolated average precision for SD/SQ and SD/TQ based on the transcriptions obtained by applying the content-based syllable bi-gram language model adapted from the baseline language model using the transcriptions of the spoken documents to speech recognition is 35.47% and 46.59%, respectively, while it is 31.00% and 45.50% based on the transcriptions obtained by applying the baseline language model to speech recognition. Altogether, the experimental results indicate that the proposed approach of applying the content-based language models to speech recognition not only improves the recognition accuracies of both the spoken documents and the speech queries but also improves the retrieval performance.

6 Conclusions

In this paper, we have proposed a novel approach, which applied the content-based language models to recognition of the spoken documents and the speech queries, to spoken document retrieval. We have tested the proposed approach on an example task for retrieval of Mandarin broadcast news. The content-based language models can be either trained with the automatic transcriptions of the spoken documents or adapted from the baseline language models using the automatic transcriptions of the spoken documents. Both kinds of content-based language models are very useful for improving the recognition accuracies of the spoken documents and the speech queries. As a result, the retrieval performance can be improved correspondingly.

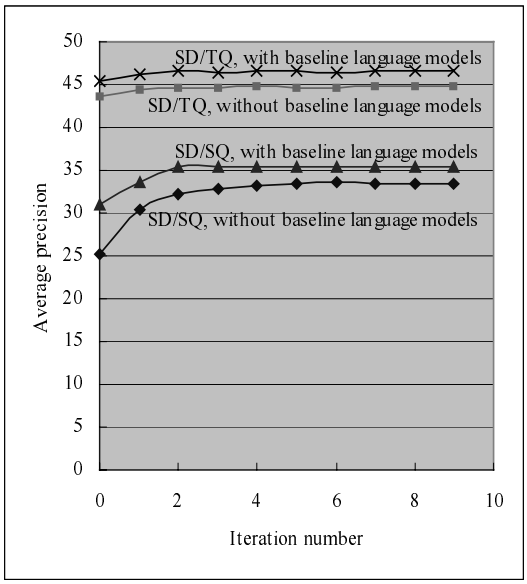


Figure 3: The performance for spoken document retrieval using speech queries (SD/SQ) and text queries (SD/TQ), with/without the baseline language models.

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 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.

11 Harman, D. Overview of the fourth text retrieval conference (TREC-4). In Proceedings of the Fourth Text REtrieval Conference, 1995.