Spoken Document Retrieval for the Languages of Hong Kong

Helen M. Meng and Pui Yu Hui
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
hmmeng@se.cuhk.edu.hk, pyhui@se.cuhk.edu.hk

ABSTRACT

The advent of the information age has brought massive digital libraries of multimedia and multilingual content. This creates a high demand for multimedia and multilingual indexing and retrieval technologies, e.g. those applicable to audio archives. This paper reports on our development of spoken document retrieval (SDR) technologies, where speech recognition is combined with information retrieval for searching audio information in Cantonese, Mandarin and English — the languages of Hong Kong. We tackle the SDR problem in a monolingual setting, where the textual queries and the audio documents are expressed in the same language. In this paper, we describe our system, our interface design for audio information visualization, as well as our experiments on spoken document retrieval.

1. INTRODUCTION

The advent of the information age has brought massive digital libraries of multimedia and multilingual content. This creates a high demand for multimedia and multilingual indexing and retrieval technologies, e.g. those applicable to audio archives. This paper reports on our work in the development of spoken document retrieval (SDR) technologies, where speech recognition is combined with information retrieval for searching audio information in Cantonese, Mandarin and English – the languages of Hong Kong. We tackle the SDR problem in a monolingual setting, where the textual queries and the audio documents are expressed in the same language.

Our initial attempt in Cantonese spoken document retrieval has been reported in [Meng et al., 2000]. Other previous work in this area include: Mandarin spoken document retrieval by [Chien et al., 1999] and [Wang et al., 1999]; and the Informedia Project from CMU [Wactlar et al., 1996] which indexed audio tracks of news broadcasts with large-vocabulary speech recognition and demonstrated multilingual capability in handling English and Serbo-Croatian.

2. EXPERIMENTAL CORPORA

2.1 Cantonese Corpus

Our spoken documents are derived from a video archive from the Hong Kong Television Broadcasts Limited (TVB). It consists of Cantonese news broadcasts from the TVB Jade[1] channel (the Cantonese channel). Table 1 shows the details about our television news corpus.

Each video clip is a news story in RealMedia file format. Each story is accompanied by a brief textual summary with a title. However, the summary is not a verbatim transcription of the audio track of the video file. Table 2 shows an example of the textual summary of a news story, together with its title (underlined).

Language	Cantonese Chinese
Source	TVB Jade channel
Number of Stories	2316 (~55.97 hours)
Extraction Period	22 June 1997 to 28 February 1998
Average Length of News	1 min 27 sec (per story)
Minimum Length of News	11 sec
Maximum Length of News	23 min 39.3 sec
Digital Video Format	RealMedia

Table 1. Information about the Cantonese news

赤柱中巴和小巴相撞十九人傷

本港今日先後發生幾宗涉及三間巴士公司，中巴、城巴和九巴的交通意外，總共造成近四十人受傷，其中最嚴重的一宗發生在赤柱馬坑村，一輛中巴和一輛滿載乘客的小巴迎頭相撞，有十九人受傷。

Table 2. An example of the textual summary of a news story. The summary title is underlined.

Very often, a news story begins with a report from the anchor(s) in the studio, followed by a live report from the field. The anchor reports are primarily studio-quality recordings. Live reports are mainly spontaneous speech (e.g. interviews) with occasional language switching (among Cantonese, Mandarin and English). The acoustic conditions in the field are highly variable, and may contain the reporter's voice-over, singing, music, applause, severe ambient noises, speaker changes, etc. These are harsh conditions for reliable automatic speech recognition.

2.2 Mandarin Corpus

We have used the collection of the 1997 DARPA HUB-4 Mandarin Benchmark from the LDC (Linguistic Data Consortium)[2] as the source of

[1] http://www.tvb.com.hk/news
[2] http://www.ldc.upenn.edu

Mandarin data. This collection includes materials, audio news and their associated machine transcriptions, that have been recorded from broadcasts by the source of CCTV (China Central Television).[3] The details of the news data are listed in Table 3.

Language	Mandarin Chinese
Source	CCTV (licensed from LDC)
Number of Stories	445 (~9.39 hours)
Extraction Period	14 January 1997 to 21 April 1997
Average Length of News	1 min 16 sec (per story)
Minimum Length of News	6.2 sec
Maximum Length of News	12 min 29 sec
Digital Audio Format	RAW

Table 3. Information about the Mandarin news

2.3 English Corpus

For English news, the TDT2 Corpus (1998 Topic Detection and Tracking project Phase 2) from LDC is used. The corpus includes the automatic transcriptions provided by Dragon recognizer[4] and the audio sources recorded from CNN (Cable Network News).[5] Table 4 shows the details of the English news data.

Language	English
Source	CNN (licensed from LDC)
Number of Stories	3414 (~43.62 hours)
Extraction Period	4 January 1998 to 30 June 1998
Average Length of News	46 sec (per story)
Minimum Length of News	7 sec
Maximum Length of News	4 min 14 sec
Digital Audio Format	RAW

Table 4. Information about the English news

3. SYSTEM OVERVIEW & CONTROL FLOW

Figure 1 shows our net-centric system architecture, with a browser-based client and various back-end servers for handling audio indexing by speech recognition, storage of the indexed audio in databases, as well as information retrieval upon the user's requests.

User can type textual queries (in Chinese or English) into the browser-based interface. These are typically keywords and key terms. The retrieval engine performs input error checks, e.g. rejecting empty and numeric inputs. Thereafter, we used word-based retrieval for English, since the English audio was pre-indexed by the Dragon recognizer. For Chinese, we map the textual query into its syllable pronunciation referencing the pronunciation dictionary CULEX,[6] and then retrieval proceeds in syllable-bigram space. The retrieval engine is pre-set to return 15 retrieved documents in real-time, and the list of documents is rank ordered by the hypothesized degree of relevance, which is described later. The user can choose to play the news clips online and/or download them to his computer.

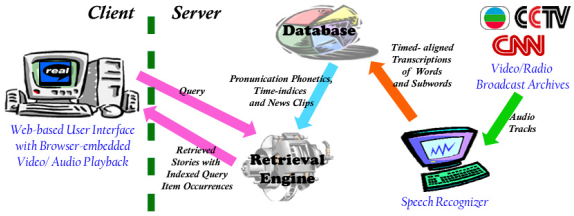

Figure 1. The diagram of the system flow

4. INTERFACE DESIGN

4.1 Information Visualization

To ease information visualization in the audio space, we designed the interface as shown in Figure 2. The entire audio track (for a news story) is depicted as a timeline, which is time-aligned with the audio track, and occurrences of the query term are indicated by pink arrows. The user can click on the arrow, at which time our system will play a 20-second window of audio centered at the position of the arrow.

Alternatively, the user may vary the window duration (of the audio segment to be played) by specification in the text boxes of the interface; or by providing start and end times of the audio segment to be extracted.

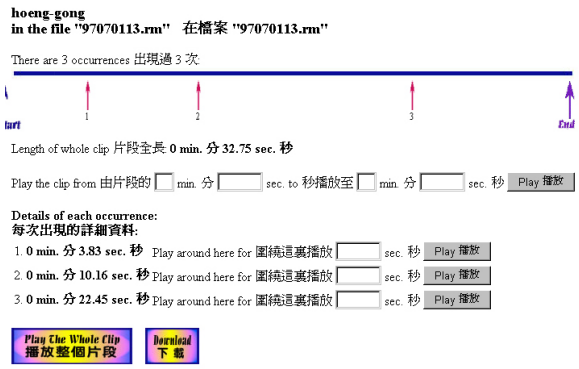

Figure 2. A sample for the web page returned by the system for the query "香港" (hoeng-gong) in the news story 97070113.rm

4.2 Incremental Search Refinement

The feature of incremental search refinement allows users to specify an additional keyword / key term, to search for news stories within the results of the previous search. The actual implementation has two retrieval steps, one for each key term, hence producing two retrieved document lists. We then extract the subset of documents present in both lists to produce the resultant retrieval list. This is equivalent to a logical "AND" applied to both key terms in retrieval. Figure 3 illustrates a scenario for the incremental search

[3] http://www.cctv.com/news
[4] http://www.dragonsys.com
[5] http://www.cnn.com
[6] http://www.ee.cuhk.edu.hk/dsp

refinement for the query "Government" after a search for the query "China". The resulting list of documents contains both “Government” and “China”.

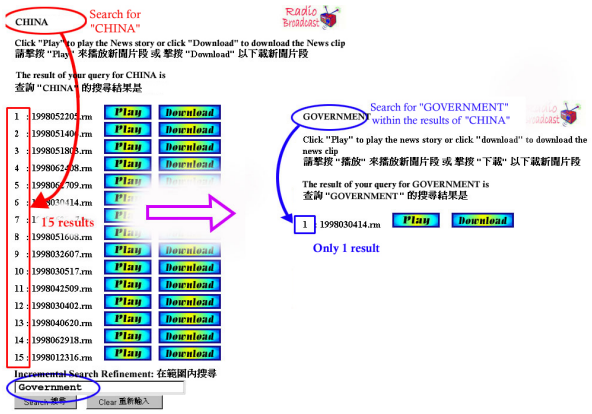

Figure 3. The result of the incremental search refinement for the query "Government" after a search for the query "China"

5. EXPERIMENTS IN CANTONESE SPOKEN DOCUMENT RETRIEVAL

We have developed our own Cantonese syllable recognizer, and evaluated both recognition and retrieval performance accuracies based on our Cantonese television news programs.

5.1 Audio Indexing

We extracted the audio tracks from the Cantonese video clips, and converted them to RealAudio format. Then, we ran our Cantonese syllable recognizer on the audio data. Since we do not have transcribed RealAudio data for training a speech recognizer, we adapted from existing resources. Our Cantonese syllable recognizer is HMM-based, and uses acoustic models based on syllable initials (I) and finals (F). The syllable initial consists of an optional onset consonant, and the syllable final consists of the vowel / diphthong followed by an optional coda consonant. We only perform base syllable[7] recognition in this work.

The recognizer was trained with phonetically-rich, continuous speech from the CUSENT corpus [Lo et al., 1998], which is recorded from a sound-proof room a microphone. Compared to the news audio tracks, which we need to index, the CUSENT recordings are from a much cleaner acoustic environment. This acoustic mismatch jeopardizes recognition performance. Therefore, we encoded the CUSENT corpus using the 8.5 kbps CELP-based speech codec of RealAudio format. This degraded data is subsequently used for training seed acoustic models. The acoustic models are context-dependent continuous density HMMs, with 16 Gaussian mixtures. We have also transcribed a small amount of original RealAudio data, and used 1.75 hours for further training, and 0.5 hours for testing. The re-training procedure was described in detail in [Meng et al., 2000].

Evaluation based on a 0.5 hour test set gave a syllable error rate of 73.8% using a syllable bigram language model. This reflects the harsh acoustic environments present in the broadcast news data, which consist of speeches from anchors (in the studio), reporters / interviewees (in the field), speech in languages other than Cantonese, and other types of non-speech sounds from live coverage of events.

To examine the performance of our retrained recognizer, we spot-checked two randomly selected news stories. Results are shown in Table 5.

	Anchor	Reporters	Interviewee
Story 1	64.3%	66.0%	89.0%
Story 2	54.9%	64.9%	100% (very noisy)

Table 5. Syllable recognition error rates of the retrained recognizer on two news stories. Performance measurements are displayed for different acoustic conditions.

The audio tracks are indexed by running our recognizer with a single-pass Viterbi to produce the recognized base syllable sequences.

5.2 Speech Retrieval

Since the stories in our corpus are not classified into topics and no relevance judgments are provided, we formulated a known-item retrieval task for our speech retrieval experiments. Recall that each audio document in our corpus as a corresponding textual summary with a title. Each summary title is used as a query to retrieve its corresponding textual or audio document from the pool. Retrieval is based on the vector-space model in SMART [Salton & McGill, 1983].

We adopted the following term weighing strategies for retrieval:

- For term i in document d:

$$d[i] = \left(0.5 + 0.5 \times \frac{tf_d[i]}{\max_i(tf_d[i])}\right) \times \ln\left(\frac{N+1}{n_i}\right)$$

- For term i in query q:

$$q[i] = \left(0.5 + 0.5 \times \frac{tf_q[i]}{\max_i(tf_q[i])}\right) \times \ln\left(\frac{N+1}{n_i}\right)$$

where $tf[i]$ is the frequency of term i in query q

N is the total number of documents, and

N_i is the number of documents with term i

The 0.5 in the above equations augments the relative $tf[i]$ value. The similarity $S(q, d)$ between a query q and document d is measured by the normalized inner product, to form the basis of retrieval as:

$$S(q,d) = \frac{q \cdot d}{\|q\| \cdot \|d\|}$$

The retrieval engine produces a list of 15 retrieved

[7] The base syllable does not contain any tone information.

documents for each query, and these are ranked according to the query-document similarity scores. The rank of the correct document, averaged over all queries, is used as our evaluation metric. The average inverse rank (AIR) is defined as:

$$AIR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$$

where N is the total number of news stories (N=2316), and $rank_i$ is the rank of relevant document in the retrieved list for query i

Our previous experiments in Cantonese spoken document retrieval [Meng et al., 2000] have shown that overlapping character / syllable bigrams are effective indexing / retrieval units. They can resolve ambiguities in Chinese word tokenization and Chinese homophones, which are problematic for speech retrieval. Hence we adopt these units for our current retrieval task. In addition, we also augment the syllable bigrams with skipped syllable bigrams, as they can capture Chinese abbreviations and helped improve retrieval performance in [Li et al., 2000]. Table 6 shows the retrieval performances based on AIR, for a variety of indexing techniques — Chinese character or base syllable bigrams and skipped bigrams.

	Bigrams	Skipped Bigrams
Text	0.834	0.818
Text converted Syllable	0.830	0.818
Recognized Syllable	0.477	0.481

Table 6. Retrieval performances based on average inverse rank, for a variety of indexing techniques – Chinese character or base syllable bigrams and skipped bigrams.

Referring to Table 6, the results for the row "Text" refers to character-based retrieval, where queries and documents are represented by overlapping character bigrams. The "Text-converted Syllables" refer to the transformation of Chinese characters into syllables by pronunciation lookup, and retrieval is based on overlapping syllable bigrams. This result approximates that of perfect speech recognition. The last row labeled "Recognized Syllable" utilized syllables output from our speech recognizer, and hence contain recognition errors. These results indicate that speech recognition errors degrade retrieval performances, which is partially salvaged by the use of skipped bigrams in indexing. An AIR value of approximately 0.5 implies that the correct document is ranked second on average in the retrieved list.

The same methodology is used for retrieval of Mandarin television news programs from CCTV. Retrieval of English audio is based on words only. We were able to port our SDR system across languages.

6. CONCLUSIONS AND FUTURE WORK

This paper describes the design and development of a spoken document retrieval system, which can handle cross-media retrieval, i.e. a textual query can retrieve audio documents. We have developed our own Cantonese recognizer for indexing Cantonese news audio, and a syllable-based retrieval engine for speech retrieval of Cantonese television news programs. We have also applied the similar retrieval techniques on Mandarin and English news programs.

In the future, we plan to enhance speech retrieval performance by incorporating a video parsing technique. The video frames provide a valuable source of information, which allow us to detect the studio-to-field transitions effectively. Video parsing is possible because anchor shots in the studio are fairly homogeneous, but live shots from the field are highly dynamic. The video parsing technique enables us to segment the audio track into an initial portion of anchor speech and subsequent portion of field speech. The cleaner acoustic environment for the anchor speech recordings imply better recognition performance, hence more reliable audio indexing for speech retrieval.

7. ACKNOWLEDGMENTS

We would like to thank the Television Broadcasts Limited for providing the Cantonese news video in this project. The first author also thanks her undergraduate students, Dias Cheung, Florence Chan and Agnes Kwok for various implementational assistance in this work. We are also grateful to Wai-Kit Lo and Yuk-Chi Li, for their help with audio indexing and evaluation of information retrieval.

8. REFERENCES

[1] Chien, L. F. and H. M. Wang, "Exploration of Spoken Access for Chinese Textual Speech Information Retrieval", Proceedings of the International Symposium on Signal Proceedings and Intelligent Systems, 1999.

[2] Li, Y. C., W. K. Lo, H. Meng and P. C. Ching, "Query Expansion using Phonetic Confusions for Chinese Spoken Document Retrieval", Proceedings of IRAL, Hong Kong, 2000.

[3] Meng, H., W. K. Lo, Y. C. Li and P. C. Ching, "Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval", Proc. of ICSLP, Beijing, 2000.

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

[5] Wactlar, H., T. Kanade, M. Smith and S. Stevens, "Intelligent Access to Digital Video: Informedia Project", IEEE Computer, Theme Issue on Digital Library Initiative, May 1996.

[6] Wang, H. M., "Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Marching", Proceedings of IRAL, 1999.