Two Robust Methods for Cantonese Spoken Document Retrieval

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

ABSTRACT

This paper reports on two methods aimed at achieving robustness for Cantonese spoken document retrieval. Our experimental corpus contains 60 hours of Cantonese television news broadcasts with over 1600 news stories. These spoken documents are indexed by automatic speech recognition of Cantonese base syllables. Recognition performance degrades significantly as we migrate from anchor speech recorded in the studio to reporter/interviewee speech recorded in the field. Recognition errors affect retrieval performance. We devised two robust methods to reduce the adverse effects of speech recognition errors on retrieval: (1) developing techniques to automatically extract studio speech from the audio tracks and using only these in retrieval; and (2) using *N*-best recognition hypotheses for document expansion prior to retrieval. Results indicate that (i) the best method to automatically extract studio speech segments fuses audio-based segmentation with video-based segmentation; (ii) using only the studio speech segments for our known-item retrieval task may not necessarily bring about better retrieval performance since we are discarding approximately three quarters of the audio in our corpus; (iii) the use of *N*-best recognition hypothesis for document expansion can bring about further improvements in retrieval performance, attaining an average inverse rank of 0.654.

1. INTRODUCTION

The exponential growth of the Internet presents a rich source of online information in a variety of media – text, audio and video. This creates a demand for technologies that can efficiently retrieve and manage multimedia information. Previous work in this area includes Mandarin[1] spoken document retrieval [1, 2] and the CMU Informedia project [3] that uses image and audio information concurrently for digital video access. We have been working on Cantonese spoken document retrieval based on local television news broadcasts [4]. Cantonese is a major dialect of Chinese, predominant in Hong Kong, Macau, South China and many overseas Chinese communities. Cantonese is monosyllabic in nature and contains between six to nine tones. We combine speech recognition and information retrieval techniques to achieve spoken document retrieval. Investigation shows that anchor speech recorded in the studio have significantly higher recognition accuracies than the reporter/interviewee speech recorded in the field. This motivates us to devise automatic methods to extract anchor/studio speech that can be reliably indexed for retrieval. We present three methods that locate studio-to-field segment boundaries prior to studio speech extraction: (i) video-based segmentation; (ii) audio-based segmentation and (iii) fusion of video- and audio-based segmentation. We have also used the *N*-best hypotheses generated by our speech recognizer for document expansion, so as to enrich the document representations with the aim to improve retrieval performance.

2. EXPERIMENTAL CORPUS

The corpus used in our spoken document retrieval experiments is derived from the Cantonese news broadcasts of the Hong Kong Television Broadcasts Limited (TVB). The corpus covers a period of four months. Table 1 shows the details of the video corpus used in this work. Each video file in the corpus corresponds to a news story and is accompanied by a text file storing a textual summary with a title. The textual summary is brief, and is by no means a verbatim transcription of the audio track of the story. On average, the length of a textual summary is approximately one fourth of its corresponding audio track if we compare in terms of the number of syllables/characters.[2] The title has an average length of 17.5 characters. Table 2 provides an example of a textual summary and its title.

Language	Cantonese Chinese
Source	TVB Jade channel
Digital Video Format	MPEG-1
Number of Stories	1627
Total Duration	~60.4 hours
Extraction Period	7 July to 17 Aug, 1999 5 Oct to 31 Dec, 2000
Average Length of News	2 min 14.6 sec (per story)
Length of news stories	4.5 sec to 8 min 55 sec

Table 1. Detailed information of the Cantonese video corpus used in our experiments.

預科生可更改報讀學科優先次序
高級程度會考昨天放榜後，預科生由今天起一連兩天，可以因應自己的成績，到大學聯招處更改報讀大學學科的優先次序。

Table 2. An example of the textual summary of a news story together with its title (underlined).

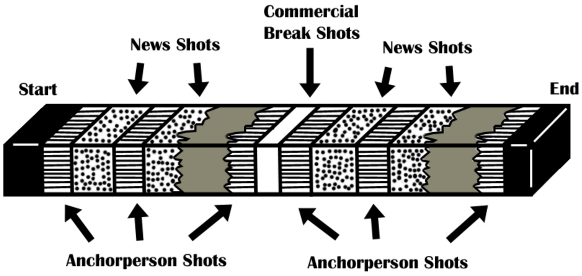


Figure 1. The temporal structure of a news program.

The news stories typically begin with a report from the anchor(s) in the studio and are optionally subsequented by a live report from the field, which is illustrated in Figure 1. We have manually labeled each news story with a segment boundary that

[1] Mandarin is the official dialect of Chinese.

[2] Each Chinese character is pronounced as a syllable.

indicates studio-to-field transition. Annotation is based *only* on the video frames.

3. AUDIO INDEXING BY AUTOMATIC SPEECH RECOGNITION

We extract the audio tracks from the video files and convert them to 16kHz monaural format. The audio is then indexed by a Cantonese *base syllable* recognizer.[3] Acoustic modeling includes three-state HMMs for syllable initials and five-state HMMs for syllable finals. These acoustic models are right-context dependent and use 16 Gaussian mixtures. Each acoustic feature vector includes 12 MFCCs with log energy and their first and second derivatives (hence 39 dimensions in total). Further details can be found in [4] and [5].

We have hand-transcribed about 2.75 hours of audio data for evaluating the performance of the speech recognizer. Syllable accuracy is 44.4%. The poor performance is due to harsh acoustic conditions (especially for field speech) and diverse speaking styles (e.g. clearly articulated speech from the anchor(s) and spontaneous speech from the reporters and interviewees).

To gauge the performance differences across various speaking styles and ambient conditions, we studied 20 news stories (a subset of the 2.75 hours mentioned above) in detail. Syllable accuracies for the anchor, reporter and interviewee speech segments are shown in Table 3. We observed severe degradation in recognition performance as we move from anchor speech recorded in the studio towards reporter/interviewee speech recorded in the field. This motivated us to use only studio speech for retrieval. Studio speech amounts to approximately a quarter of the audio tracks (in terms of duration) in our corpus. Hence using only studio speech for retrieval may reduce the computational costs required in expensive procedures such as audio indexing and may potentially give better retrieval performance.

	Studio	**Field**	
	Anchor	**Reporter**	**Interviewee**
Syllable Accuracies	**59.3%**	**43.3%**	**27.0%**
		39.2% (overall)	

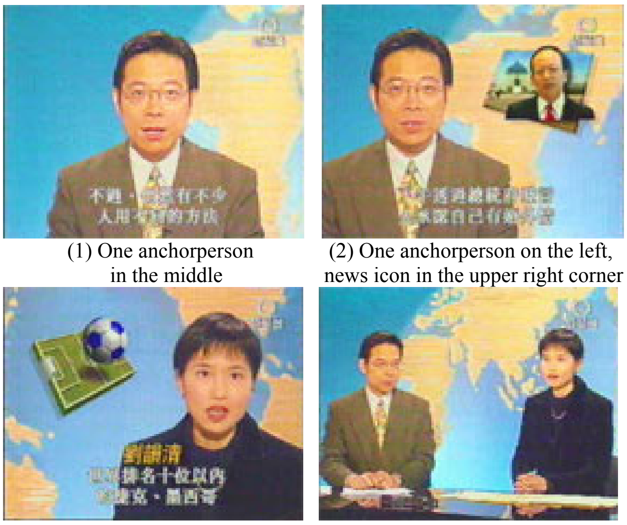
Table 3. Speech recognition performance (syllable accuracies) for anchor speech recorded in the studio (i.e. clearly articulated speech from favorable ambient conditions) and report/interview speech recorded in the field (i.e. spontaneous speech recorded from possibly harsh acoustic conditions).

4. AUTOMATIC EXTRACTION OF ANCHOR/STUDIO SPEECH SEGMENTS

We have devised three automatic methods for extracting anchor/studio speech segments. The first method utilizes video frame information only; the second utilizes audio information only and the third method fuses both audio and video information for extraction.

4.1 Video-based Segmentation

Our video-based segmentation algorithm takes advantage of the relative homogeneity of the anchor shots in the studio in comparison with the dynamically changing shots from the field. We compute the differences between adjacent video frames in terms of the spatial difference metric and color difference metric. A fuzzy c-means algorithm is used to detect adjacent frame pairs with significant changes and these are labeled as shot boundaries. The first frame is used as a key frame to represent each shot in between boundaries. Classification of the key frames via a graph-theoretic clustering algorithm yields four types of anchor shots – (i) anchor in the center; (ii) anchor on the left with an icon in the right; (iii) anchor on the right with icon on the left as well as (iv) two anchors side by side (see Figure 2). These video segments are extracted and their audio tracks are identified as studio quality anchor speech. Details of the algorithm can be found in [6].



(1) One anchorperson in the middle

(2) One anchorperson on the left, news icon in the upper right corner

(3) One anchorperson on the right, news icon in the upper left corner

(4) Two anchorpersons side by side

Figure 2. The four typical patterns of anchor shots in our video corpus.

This video segmentation algorithm is applied to all the 1627 new stories in our corpus and evaluated against the hand-labeled reference boundaries. Annotators marked studio-to-field transitions based *only* on shot changes in the video frames. The manual annotations indicate that 1545 of the news stories (~95.0%) contain a single studio-to-field transition while the remaining news stories have no field shots. A studio-to-field segment boundary automatically labeled by our video-based algorithm is considered correct if it lies within two seconds (i.e. 50 video frames) of the manually labeled segment boundary. Our video-based extraction algorithm correctly identified 1365 of the anchor/studio speech segments, achieving precision and recall values of 0.954 and 0.884 respectively (see Table 4).

4.2. Audio-based Segmentation

We made another attempt to extract anchor/studio segments based only on the audio information. Our method aims to capture differences in the acoustic signal since studio speech tends to have little noise, while field speech may contain music, environmental noises, etc. We use single-state Gaussian Mixture Models (GMM) [7, 8] for the audio-based segmentation. We trained one GMM to be the *studio model* and another to be the *field model* by applying the Baum-Welch algorithm on 5 hours of audio data (a subset of the 60.4 hours mentioned in Table 1) from our corpus. The number of Gaussian mixtures was increased exponentially from 1 to 64 during the training stage. At 64 mixtures, the GMMs can correctly extract most of the anchor/studio speech segments from the 5 hours of training data.

We used these GMM models to process the entire audio data set (60.4 hours), in order to distinguish news stories with no field shots from those with studio-to-field transitions. Hence for a news story with T speech frames, we first compute the cumulative score by traversing with the studio model only:

[3] The base syllable does not include tone information.

$$Score_{studio\_only} = \prod_{t=1}^{T} \sum_{i=1}^{64} w_i \cdot N_{studio}\left(\mu_i, \sigma_i\right) \qquad (1)$$

where w_i are the weights for the Gaussians and $N_{studio}(\mu, \sigma)$ from the studio model

Then we concatenate the studio and field models and traversed the T speech frames with a single-pass Viterbi algorithm to compute:

$$Score_{studio\_to\_field} =$$

$$\prod_{t=1}^{T_t} \sum_{i=1}^{64} w_i N_{studio}\left(\mu_i, \sigma_i\right) \prod_{t=T_t}^{T} \sum_{j=1}^{64} w_j N_{field}\left(\mu_j, \sigma_j\right) \qquad (2)$$

If $Score_{studio\_only} < Score_{studio\_to\_field}$, our audio-based segmentation framework assumes that there is a studio-to-field transition at frame T_t. Otherwise we assume that the news story consists entirely of studio speech.

We evaluate this audio-based segmentation algorithm with reference to the manually labeled studio-to-field segment boundaries. Evaluation allows a two-second deviation, similar to reported results from video-based segmentation. Results are shown in Table 4 together with the video-based segmentation algorithm. It should be noted that manual annotation is based on video frames and we have found 306 news stories (~20% of the entire corpus) in which the video scene changes from studio to field yet the anchor continues to speak until the end of the story. Hence our evaluation method may *over-penalize* the audio-based segmentation algorithm. This is reflected in the larger deviations between the automatically located boundaries and the manually labeled boundaries (see Table 4). We studied with greater care the 306 news stories where studio-to-field transitions occur only in the video but not the audio track and compared them with the 251 news stories that our audio-based segmentation algorithm claimed had no transitions. We found that 192 news stories were labeled correctly, which corresponds to a precision of 0.765 and recall of 0.627 within this special subset of news stories. Table 4 summarizes the performance of the video- and audio-based segmentation algorithms.

	Video-based	Audio-based
Number of transitions labeled by algorithm	1431	1376
Number of transitions labeled ***correctly*** (dev. less than 2 sec)	1365	1208
Precision	0.954	0.878
Recall	0.884	0.743
Mean deviation from reference boundary	0.0036 sec	-1.37 sec
Standard deviation	11.9 sec	18.8 sec

Table 4. Automatic location of studio-to-field transition boundaries by means of two methods – the first uses video information only and the second uses audio information only. Manually annotated reference boundaries were labeled based on the video frames, hence evaluation may over-penalize the audio-based segmentation method. The negative sign of the mean deviation in audio-based segmentation indicates that reporters usually start to speak a second after the video scene changes.

4.3 Fusion of Video- and Audio-based Segmentation

We attempt to fuse results from video-based segmentation with those from audio-based segmentation to further improve automatic location of studio-to-field transitions. Table 5 shows statistics relating to the presence/absence of studio-to-field transitions in the audio and video tracks of our news stories:

	Transition in audio	No transition in audio
Transition in video	**1239**	**306**
No transition in video	**0**	**82**

Table 5. Number of news stories in our corpus with presence/absence of studio-to-field transitions in the audio/video tracks. The total number of news stories is 1627.

Based on these statistics we devise the following *fusion strategy*:

Case 1: *Both* video- and audio-based segmentation algorithms detect studio-to-field transitions – we extract the anchor/studio segment according to the video-based algorithm, since its boundaries deviate less from the reference boundaries (see Table 4).

Case 2: *Only* the video-based segmentation algorithm detects a studio-to-field transition – we use the entire audio track for retrieval since there exists news stories in this category (see Table 5).

Case 3: *Only* the audio-based segmentation detects a studio-to-field transition – the entire audio track is used in spoken document retrieval since no such news story should exist (see Table 5).[4]

Case 4: *Both* video- and audio-based segmentation do not detect any transition – the entire audio track is used in spoken document retrieval.

5. SPOKEN DOCUMENT RETRIEVAL EXPERIMENTS

5.1 Retrieval models and indexing units

In our experiments, retrieval is based on the vector-space model in SMART [9] and details were described in [4]. The document term weighing (3) and query term weighing (4) used in our retrieval experiments are:

$$d[i] = \ln(tf_d[i]) + 1.0 \qquad (3)$$

$$q[i] = [\ln(tf_q[i]) + 1.0] \times \ln(\frac{N_d + 1}{n_i}) \qquad (4)$$

where $tf_d[i]$ is the frequency of term i in document d
$tf_q[i]$ is the frequency of term i in query q
N_d is the number of documents
n_i is the number of documents with term i

Each query/document is represented as a vector of indexing terms. In this work, syllable bigrams and skipped bigrams are used because this representation has previously been shown to give the best retrieval performance [4]. Figure 3 illustrates the process of forming such a representation from the textual query or document (i.e. the summary). Character bigrams/skipped bigrams are first formed from the textual word and these are then converted into syllable bigrams/skipped

[4] None of the stories fall into Case 3.

bigrams by pronunciation lookup. For audio documents, the syllables output from recognition during indexing are directly used to generate the syllable bigrams/skipped bigrams.

word:	中文大學 /zung man daai hok/
character bigrams:	中_文 文_大 大_學
syllable bigrams:	/zung_man/ /man_daai/ /daai_hok/
skipped character bigrams:	中_大 文_學
skipped syllable bigrams:	/zung_daai/ /man_hok/

Figure 3. Procedure for forming text-converted syllable bigrams/skipped bigrams. The syllable labels follow the Cantonese LSHK[5] convention.

5.2 Known-Item Retrieval task

We have formulated a *known-item retrieval task* (KIR) based on our video corpus. The summary title of each news story is used as a query to retrieve its corresponding textual document (i.e. the summary) or audio document (i.e. the audio track) from the archive.

Since there is only one relevant textual/audio document for each query, we adopt the average inverse rank (AIR) as our evaluation criterion:

$$AIR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \qquad (5)$$

where N is the total number of news stories (N=1627) and $rank_i$ is the rank of relevant document in the retrieved list for query i

Perfect retrieval will produce AIR=1, while poor retrieval will give small values for AIR.

5.3 Experimental Results

Retrieval based on text-converted syllable bigrams and skipped bigrams (converted from the textual summaries) provided an approximate benchmark for the case of perfect syllable recognition, with AIR=0.971. Retrieval based on the indexed spoken documents (i.e. using the audio tracks) gave lower performance due to imperfect syllable recognition. Table 6 shows the retrieval results for various methods of extracting the anchor/studio speech segments.

Method for locating studio-to-field transition (and extraction of studio speech segment)	**AIR**
No extraction (entire audio track is used)	0.633
Automatic video-based segmentation	0.628
Automatic audio-based segmentation	0.631
Fusion of video- and audio-based segmentation	**0.641**

Table 6. Spoken document retrieval performance based on extracted anchor/studio speech segments. The result without using any extraction method (entire audio tracks is used) is included as a reference (shaded). Fusion of video- and audio-based segmentation gives the best retrieval result.

Results suggest that using only the studio speech segments in retrieval may not necessarily improve retrieval performance over the baseline (i.e. when the entire audio track is used). A possible reason is that we are discarding approximately three quarters of the audio in our corpus. Audio-based segmentation improved slightly over video-based segmentation since it can correctly handle news stories for which the studio-to-field transitions occur in the video but not the audio. Fusion of video- and audio-based segmentation gave the best performance.

6. DOCUMENT EXPANSION USING *N*-BEST RECOGNITION HYPOTHESES

In this work, we attempt to apply a document expansion technique [10] based on *N*-best recognition hypotheses. This method aims to enrich the document representations and reduce the adverse effect of speech recognition errors on retrieval performance. For a retrieval task involving textual queries and spoken documents, the textual queries need to be mapped into base syllables by pronunciation dictionary lookup, and the spoken documents are transformed into a syllable-based representation via speech recognition. Hence the query syllables are error-free while the document syllables are errorful. Document expansion using *N*-best recognition hypotheses may help to bridge this gap between queries and documents.

6.1 *N*-best Hypothesis from Recognition

N-best recognition hypotheses used in this work are the *N* most probable syllable sequences output from the Cantonese syllable recognizer (N = 5 in this work). Differences among these *N*-best recognition hypotheses may occur only in a few syllables. An example extracted from our experiment is shown in Table 7.

```
Filename: "199907070101.rec"

1: ... jik wui sei nang ...
2: ... jik wui sei nang ...
3: ... jik wui zau nang ...
4: ... jik wui sei nang ...
5: ... jik wui zau nang ...
```

Table 7. An example of the *N*-best syllable sequences output from the recognizer. It can be seen that within the four-syllable window as shown, /sei/ has been misrecognized as /zau/ in two of the five recognition outputs.

6.2 Re-weighing Different Retrieval Units

Syllables that appear consistently across the *N*-best recognition outputs are likely to be more reliable and hence should be weighed more heavily in the document vector.

As mentioned earlier, every document is represented as a vector of syllable bigrams and skipped bigrams. We formed the bigrams from the *N*-best recognition hypotheses as shown in Table 8 and re-weighed them as illustrated in Table 9. We use the number of occurrences of each token in all five hypotheses to be the weight of that token for retrieval. Since we have five hypotheses, the maximum weight of a token is five.

```
"199907070101.rec"

... jik_wui wui_sei sei_nang jik_sei wui_nang ...
... jik_wui wui_sei sei_nang jik_sei wui_nang ...
... jik_wui wui_zau zau_nang jik_zau wui_nang ...
... jik_wui wui_sei sei_nang jik_sei wui_nang ...
... jik_wui wui_zau zau_nang jik_zau wui_nang ...
```

Table 8. An example on bigrams and skipped bigrams formed with the hypothesized syllables listed in Table 7.

[5] Linguistic Society of Hong Kong

```
"199907070101.rec"

...
jik_wui 5 wui_sei 3 wui_zau 2 sei_nang 3
zau_nang 2 jik_sei 3 jik_zau 2 wui_nang 5
...
```

Table 9. Re-weighing the different bigrams based on alternate recognition hypotheses in Table 8. Since /wui_sei/ and /wui_zau/ have occurrences of three and two respectively in the hypotheses, they have the adjusted weights of three and two. This is also the case for /sei_nang/, /zau_nang/, /jik_sei/ and /jik_zau/. Their weights are smaller than other bigrams that appear consistently across the recognition hypotheses. These bigrams have weights of five.

6.3 Retrieval Performance with Document Expansion using *N*-best Recognition Hypotheses

Document vectors are expanded according to the additional syllable bigrams and skipped bigrams derived from the *N*-best recognition hypotheses. The weighing function of the indexing terms is modified as shown in Equation (6):

$$d[i] = \ln(tw_d[i]) + 1.0 \qquad (6)$$

where $tw_d[i]$ is the weighed term frequency of term i in document d

Retrieval experiments have been performed using these expanded documents in the KIR task described before. Table 10 shows the retrieval results with document expansion using *N*-best recognition hypotheses for the anchor speech extracted using various segmentation methods.

Method for locating studio-to-field transition (and extraction of studio speech segment)	**AIR**
No extraction (entire audio track is used)	0.652
Automatic video-based segmentation	0.639
Automatic audio-based segmentation	0.650
Fusion of video- and audio-based segmentation	**0.654**

Table 10. Spoken document retrieval performance based on extracted studio speech segments and including indexing terms from *N*-best recognition hypotheses. The result without using any extraction method (entire audio track is used) is included as a reference (in shaded region). Fusion of video- and audio-based segmentation gives the best retrieval results.

Comparison between Table 10 and Table 6 suggests that the use of *N*-best recognition hypotheses for document expansion improves retrieval performance. Audio-based segmentation achieves slightly better performance than video-based segmentation and that fusion of the two segmentations results gives the best performance.

7. CONCLUSIONS AND FUTURE WORK

This paper reports on two methods aimed at achieving robustness for Cantonese spoken document retrieval. We have developed an experimental corpus that contains 60 hours of Cantonese television news broadcasts with over 1600 news stories. These spoken documents are indexed by automatic speech recognition of Cantonese base syllables. Recognition performance degrades significantly as we migrate from anchor speech recorded in the studio (at 59% syllable accuracy) to reporter/interviewee speech recorded in the field (at 39% syllable accuracy). Recognition errors affect retrieval performance. We attempt to reduce the adverse effects of speech recognition errors on retrieval by (1) developing techniques to automatically extract studio speech from the audio tracks and using only these in retrieval; and (2) using *N*-best recognition hypotheses for document expansion prior to retrieval.

We have developed three automatic methods for locating studio-to-field transitions (and hence the studio segments) in the video news stories: (i) video-based segmentation distinguishes between the more homogeneous studio shots from the more dynamic field shots; (ii) audio-based segmentation uses a Gaussian Mixture Model (GMM) to distinguish the cleaner studio recordings from the noisier field recordings; and (iii) a fusion strategy that combines video- and audio-based segmentation. Results based on a known-item retrieval task indicate that only using studio segments may not necessarily improve retrieval performance over the baseline (i.e. using all of the studio and field speech), possibly because we are discarding approximately three quarters of the audio data. Fusion of video-based segmentation and audio-based segmentation gave the best retrieval performance, achieving AIR=0.641.

In an attempt to improve retrieval performance further, we used the *N*-best recognition outputs for document expansion. The alternative recognition hypotheses introduce additional indexing terms (syllable bigrams and skipped bigrams) that are re-weighed in our vector-space retrieval model. Results show that augmenting the top-scoring recognition hypothesis with *N*-best hypotheses brought consistent improvements, achieving AIR=0.654. Future investigation will be devoted to the *selective* use of (noisy) field speech in retrieval.

ACKNOWLEDGMENTS

We thank the Television Broadcasts (HK) Limited for providing the Cantonese news video content. This work is partially supported by a grant from the Hong Kong SAR Government Research Grants Council (Project No. CUHK 4223/01E) and a grant from the Hong Kong SAR Government University Grant Council under Area of Excellence in Information Technology.

REFERENCES

1. Chien, L. F. and H. M. Wang, “Exploration of Spoken Access for Chinese Text and Speech Information Retrieval,” *Proceedings of the ISSPIS*, pp. 578-583, Guangzhou, China, 1999.
2. Wang, H. M., “Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching,” *Proceedings of IRAL*, Taipei, Taiwan, 1999.
3. Wactlar, H., T. Kanade, M. Smith and S. Stevens, “Intelligent Access to Digital Video: Informedia Project,” *IEEE Computer*, vol. 29, pp.46-52, May 1996.
4. Meng, H., W. K. Lo, Y. C. Li and P. C. Ching, “Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval,” *Proceedings of ICSLP*, pp. 101-104, Beijing, China, 2000.
5. Meng, H., X. Tang, P. Y. Hui, X. Gao and Y. C. Li, “Speech Retrieval with Video Parsing for Television News Programs,” *Proceedings of ICASSP*, pp. 1401-1404, Salt Lake City, USA, 2001.
6. Hui, P. Y., X. Tang, H. Meng, W. Lam and X. Gao, “Automatic Story Segmentation for Spoken Document Retrieval,” *Proceedings of FUZZ-IEEE,* Melbourne, Australia, 2001.
7. Chen, T., C. Huang, Eric Chang, and Jingchun Wang, “Automatic accent identification using Gaussian mixture models,” *Proceedings of the ASRU 2001*, pp. 343-346, Trento, Italy, 2001.
8. Reynolds, D., “Speaker identification and verification using Gaussian mixture speaker models,” *Speech*

Communication, vol. 17, pp. 91-108, Elsevier Science, 1995.

9. Salton, G. and M. McGill, "*Introduction to Modern Information Retrieval*," McGraw-Hill, New York, 1983.

10. Singhal, Amit and Fernando Pereira, "Document Expansion for Speech Retrieval," *Proceedings of ACM SIGIR*, pp 34-41, Berkeley, USA, 1999.