Speech Retrieval with Video Parsing for Television News Programs

Helen M. Meng[1], *Xiaoou Tang*[2], *Pui Yu Hui*[1], *Xinbo Gao*[2] *and Yuk Chi Li*[1]
[1]Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
[2]Department of Information Engineering,
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
hmmeng@se.cuhk.edu.hk

ABSTRACT

We have been working on speech retrieval from Chinese (Cantonese) television news programs. The use of automatic speech recognition for audio indexing produces imperfect transcriptions, and recognition errors affect retrieval performance. A news story typically contains a brief report by the anchor person(s) in the studio, as well as news footage from the field. Investigation shows that our recognizer performs better when indexing audio from the studio, compared to that from the field. In order to automatically extract the "reliable" audio segments for speech retrieval, we attempt to detect studio-to-field transitions by means of video parsing. Our study is based on 146 news stories collected from local television Cantonese news programs. We formulated a known-item retrieval task and adopted the average inverse rank (AIR) as our evaluation metric. Retrieval is performed based on syllable bigram units, augmented with skipped syllable bigrams. Retrieval using the entire audio track of each news story gave AIR=0.759. With the incorporation of video parsing, we performed retrieval based *only* on the studio recordings, which produced AIR=0.768.

1. INTRODUCTION

The explosive growth of the Internet has created a rich source of electronic information in a variety of media – text, audio and video. This creates a demand for multilingual and multimedia information retrieval technologies to enable the user to retrieve personally relevant content on demand. Text-based search engines are widely used, and audio / video searching are active areas of research. We have been working on the problem of Chinese spoken document retrieval [Meng et al., 2000]. In particular, we work with Cantonese, which is a major dialect of the Chinese language, commonly used in Hong Kong, Macau, South China and many overseas Chinese communities. This work attempts to apply the video parsing technique to assist our Cantonese spoken document retrieval task, based on television news programs. We combine the technologies of speech recognition and video parsing for indexing our audio tracks, and applied a vector-space model for retrieval. Previous work in this area include Mandarin (the major dialect of Chinese) spoken document retrieval by [Chien et al., 1999] and [Wang et al., 1999]; and the CMU Informedia project [Wactlar et al., 1996] which uses image and audio information concurrently for digital video access.

2. CORPORA

Video content for our experiments is provided by the Hong Kong Television Broadcasts Ltd. (TVB). It consists of Cantonese news broadcasts from the Jade[1] channel (i.e. the Cantonese channel), with 146 news stories, for which Table 1 provides some detailed information.

Language	Cantonese Chinese
Source	TVB Jade channel
Number of Stories	146 (~3.11 hours)
Extraction Period	7-9, July 1999 and 5-17, October 2000
Average Length of News	1 min 15.62 sec (per story)
Minimum Length of News	7.13 sec
Maximum Length of News	4 min 0.2 sec
Digital Video Format	MPEG-1

Table 1. Information about the video content used on our experiments.

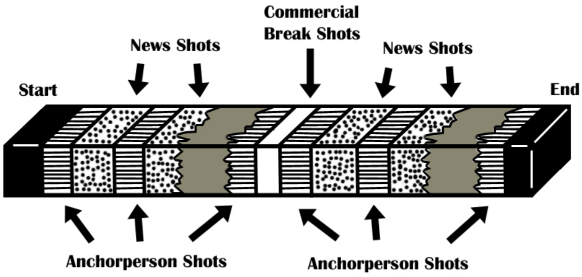


Figure 1. The temporal structure of a news program.

Each MPEG file contains a single news story manually segmented from the news program, which is illustrated in Figure 1. Each story is accompanied with a brief textual summary and its title. However, the summary is not a verbatim transcription of the audio track of the video file. We estimated that the length of the textual summary is roughly a quarter that of the audio track, measured in the number of characters / syllables.[2] The average length of the summary titles is 17.5 characters. Table 2 shows an example of the textual summary of a news story, together with its title (underlined).

立法會否決檢討行政會議委任及權責的動議

行政會議成員在公務與私人事業之間的角色衝突，近日經常引起爭論。立法會今日經過兩個小時的激烈辯論之後，議員最終否決由議員何秀蘭提出，檢討行政會議的委任以及權責的動議。

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

Very often, the 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 in Cantonese. Live reports are mainly spontaneous speech (e.g. from

[1] http://www.tvb.com.hk/news

[2] Written Chinese consists of a sequence of characters. Each character is pronounced as a syllable.

interviews) with occasional language switching (among Cantonese, Mandarin and English, from the trilingual Hong Kong environment), and recorded from highly variable acoustic conditions, e.g. with the reporter's voice–over, singing, music, applause, severe ambient noises, etc. – these are harsh conditions for reliable automatic speech recognition.

3. AUDIO INDEXING BY AUTOMATIC SPEECH RECOGNITION

We have extracted the audio tracks from the MPEG-1 video files of the news stories, and converted them to RealAudio format. This is because the acoustic models of our recognizer has previously been trained on RealAudio data. Our Chinese 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[3] recognition in this work.

Training data for speech recognition consists of:

- Cantonese continuous read speech (2.25 hours) recorded in a studio, based on the hand-transcribed CUSENT corpus [Lo et al., 1998], which is encoded with the 8.5 kbps CELP-based speech codec of RealAudio format; and
- 1.75 hours of RealAudio speech data, obtained from the TVB Cantonese news programs between 1997 and 1998,[4] and the data has been manually transcribed.

The acoustic models are context-dependent continuous density HMMs, with 16 Gaussian mixtures. They were initially trained with the clean, read speech from CUSENT, then adapted for news audio by embedded retraining with the news data. This was described in detail in [Meng et al., 2000].

Testing data for speech recognition consists of:

- 3.11 hours of RealAudio speech data (converted from MPEG-1), and the data has been manually transcribed for recognition evaluation.

Evaluation based on this data test set gave a syllable accuracy of 25.71%. This reflects the great mismatch in recording conditions and speaking styles between training and testing. The quality difference between RealAudio converted from MPEG and the “original’ RealAudio data, and the harsh acoustic environments encountered in the news audio. We spot checked the syllable accuracies of 11 audio stories, by hand-segmenting it into portions of anchor / report / interviewee speech, manually transcribing the audio tracks and then comparing with the recognizer's syllable hypotheses. Results are shown in Table 3.

Anchor	**Reporter**	**Interviewee**
34.8%	25.58%	5.16%

Table 3. Syllable accuracies for all the news stories in the test set. The audio track is hand-segmented into portions of anchor / reporter / interviewee speech. The anchor segment is recorded in the studio. The reporter and interviewee segments belong to live news footage from the field.

There is severe degradation in recognition performance as we move from studio-quality anchor speech to news reported from the field. It should be desirable to devise methods for automatically detecting studio-to-field transitions, so that we can place heavier emphasis on anchor speech for retrieval purposes. The video frames should provide a reliable source of information for detection, since video frames shot in the studio are fairly homogeneous, in contrast to those shot in the field.

4. VIDEO PARSING

Our study thus far indicates that speech recognition performance for audio indexing is far more reliable for the anchor speech than the reporter's and interviewee's speech from the field. In this section we will describe our method of video parsing, which automatically detects scene changes from the video frames in order to locate the transition from the studio to the field. With this information, we can segment the audio track for each story into the segment of anchor speech, and the segment of live report. Thereafter, these segments can be processed individually for speech retrieval.

We found that the temporal syntax of the news video from our local television station is rather straightforward. The majority of our Cantonese news stories consist of two parts – the anchor shots followed by the field shots. Also, an *entire* news program contains a series of news stories interleaving with commercial breaks (see Figure 1).

We have manually labeled the studio-to-field transitions for all the 146 stories (or video files) in our corpus. We found that approximately 91% of the stories follow the temporal syntax described above. The few remaining cases have one of the following temporal syntax: (i) anchor shots only; or (ii) anchor-field-anchor combination.

We adopted the video parsing scheme developed by [Gao & Tang][5]. It consists of four modules as shown in Figure 2. First, two metrics are used to compare the qualitative difference between two consecutive frames. They are the spatial difference metric (SDM) and the histogram difference metric (HDM).

$$SDM = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left| I_t(i,j) - I_{t+1}(i,j) \right| \qquad (1)$$

$$HDM = \frac{1}{M \times N} \sum_{k=1}^{L} \left| H_t(k) - H_{t+1}(k) \right| \qquad (2)$$

where $M \times N$ is the frame size,

$I_t(i, j)$ denotes the intensity of a pixel at location (i, j),

$H_t(k)$ denotes the number of pixels with color k in the t-th frame, and

L is the total number of colors.

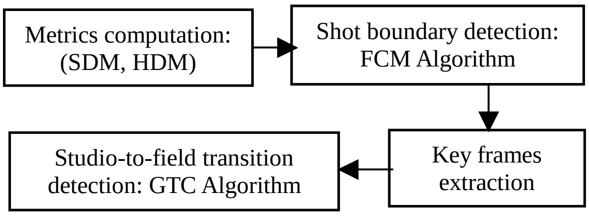

Metrics computation: (SDM, HDM)
Shot boundary detection: FCM Algorithm
Key frames extraction
Studio-to-field transition detection: GTC Algorithm

Figure 2. The framework of our video parsing scheme.

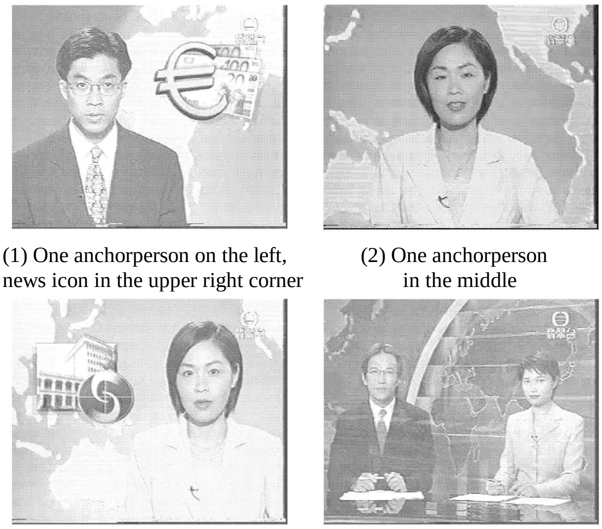
SCM and HDM both generate a sequence of pulses. A high pulse corresponds to a shot boundary. After normalizing the SDM and HDM metrics, we use the fuzzy c-means (FCM) clustering algorithm [Bezdek 1981] to detect the shot boundaries. Based on these shot boundaries, each video frame can be assigned to an individual shot. There exist two categories of shots, i.e., studio shot and field shot. To detect the studio-to-field transition, it is necessary to distinguish these two categories of shots. To this end, we first extract the key frame for each shot. In general, this process is simplified by taking one frame from each shot. Although there are several patterns of studio shots (anchor

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

[4] We have also collected video data from two years ago, which was in RealMedia format. However, the content provider has since adopted an alternate video format, which created discontinuity in our data usage.

[5] Co-authors of this paper.

frames), such as the four typical patterns shown in Figure 3, we observe that each pattern appears no less than twice in a news program. Because the background region tends to be relatively fixed, the different studio key frames with the same pattern must have a similar color histogram. Using this similarity we can group and detect studio shots of the same pattern in a self-organized fashion through the graph-theoretical cluster (GTC) method. The boundary frames between studio shots and field shots are identified as studio-to-field transitions.



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

(2) One anchorperson in the middle

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

(4) Two anchorpersons

Figure 3. The four typical patterns of anchor shots in our video corpus.

4.1 Performance on Video Parsing

We applied the video parsing algorithm on all the 146 video files in our corpus. As mentioned earlier, manual verification shows that there are 133 of these (~91%) which display the anchor-field temporal syntax (and thus contains one studio-to-field transitions).

Our video parsing algorithm labeled 112 news stories with a single studio-to-field transition. The remaining stories were labled with either zero or multiple transitions. We consider a studio-to-field transition to be "correctly detected" if the automatically labeled transition frame and the manually labeled one deviate no more than a distance of 50 frames (which corresponds to approximately 2 seconds of audio). Evaluation shows that all 112 of the automatically detected transitions correspond to the manual ones. Hence video parsing achieved a precision of 1, and a recall of 0.842.

For each news story, we can use the transition frame number to segment the audio track into two portions – the first portion corresponds to the anchor's speech (according to our temporal syntax), and the second portion corresponds to the reporter's / interviewee's speech.

5. SPEECH RETRIEVAL

5.1 A Known-Item Retrieval Task

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*:

$$d[i] = \left(0.5 + 0.5 \times \frac{tf_d[i]}{\max_i(tf_d[i])}\right) \times \ln\left(\frac{N+1}{n_i}\right) \qquad (3)$$

- For term *i* in query *q*:

$$q[i] = \left(0.5 + 0.5 \times \frac{tf_q[i]}{\max_i(tf_q[i])}\right) \times \ln\left(\frac{N+1}{n_i}\right) \qquad (4)$$

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:

$$S(q,d) = \frac{q \cdot d}{\|q\| \cdot \|d\|} \qquad (5)$$

The retrieval engine produces a list of 15 retrieved 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:

$$AIR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i} \qquad (6)$$

where *N* is the total number of news stories (*N*=146), 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,[6] as they can capture Chinese abbreviations and helped improve retrieval performance in [Li et al., 2000]. In short, we form the following query / document representations for retrieval:

(i) Overlapping character bigrams – the query vector representation contains overlapping character bigrams derived from the summary title; and the document vector representation contains overlapping character bigrams derived from the textual summary prose.

(ii) Overlapping text-converted syllable bigrams – the above character bigrams are converted into syllable bigrams (to which we refer as *text-converted* syllable bigrams), by pronunciation lookup from a Cantonese dictionary CULEX. These may be augmented with skipped syllable bigrams (from text-conversion).

(iii) Overlapping recognized syllable bigrams – the syllable bigrams for the queries are same as (ii), but the syllable bigrams for the documents are derived from syllable recognition of the audio document. These may be augmented with skipped syllable bigrams (from recognition).

[6] For the syllable sequence $s_1, s_2, s_3, s_4, \ldots$, the skipped syllable bigrams are $s_1s_3, s_2s_4, \ldots$

The character-bigrams and text-converted syllable bigrams serve as a reference upper bound for the performance in our speech retrieval task.

5.2 Experimental Results

Table 4 shows the speech retrieval performance based on character bigrams and text converted syllables. Retrieval performance is high, which is an artifact caused by the small number of stories in our known-item retrieval task. Also the results suggest that the summary titles can succinctly capture the key terms in the news story.

Retrieval Unit	**AIR**
Character Bigrams	0.990
Character Bigrams + Skipped Bigrams	0.993
Text-converted Syllable Bigrams	1
Syllable Bigrams + Skipped Bigrams	1

Table 4. Retrieval performances based on average inverse rank using character bigrams and text-converted syllable bigrams.

Retrieval Unit	**AIR**		
	Full	**Studio**	**Field**
Rec. Syllable Bigrams (Manual Video Parsing)	0.731	0.714	0.411
Rec. Syllable Bigrams + Skipped Bigrams (Manual Video Parsing)	0.759	0.765	0.390
Rec. Syllable Bigrams (Automatic Video Parsing)	0.731	0.714	0.421
Rec. Syllable Bigrams + Skipped Bigrams (Automatic Video Parsing)	0.759	0.768	0.406

Table 5. Retrieval performances based on average inverse rank using recognized syllable bigrams. We include the use of the full audio track (**Full**), speech from the studio only (**Studio**) and speech from the field only (**Field**). The shaded rows show results based on manual video parsing for studio-to-field transitions; and the unshaded rows show results based on automatic video parsing.

The results in Table 5 show retrieval performances using recognized syllable bigrams and skipped bigrams. We include results based on using the full audio tracks from the news stories; only the anchor's speech from the studio; and only the speech from the field. The shaded rows are results based on manual video parsing for studio-to-field transitions; and the unshaded rows are results based on automatic video parsing. For news stories where video parsing fails to produce a single transition point, the entire audio track is used instead. We observe an overall degradation of retrieval performance (c.f. Table 4) due to imperfect recognition. Within Table 5, anchor speech results are comparable with those from full audio, while the field speech suffers a severe degradation. Results from automatic video parsing comparable to manual video parsing.

It is encouraging to see that the retrieval performance based only on anchor speech approaches that based on the full audio. This result suggest that our video parsing is reasonably effective in helping to locate the anchor speech segments from the news audio tracks.

6. CONCLUSIONS AND FUTURE WORK

This paper reports on our preliminary attempt in fusing video and audio information for speech retrieval of Cantonese television programs. Most news stories exhibit the same temporal syntax, i.e. the anchor's report is followed by the live coverage. The performance of speech recognition varies greatly between the anchor's speech from the studio to the reporter's / interviewee's speech from the field. The former is well-articulated and recorded from a controlled acoustic environment. The latter is more spontaneous and often recorded in harsh acoustic conditions. This is a major concern since the reliability of audio indexing directly affects retrieval performance.

We attempt to enhance speech retrieval performance by incorporating the video parsing technique. The video frames provide a valuable source of information, which allow us to detect the studio-to-field transitions effectively. This is because anchor shots in the studio are fairly homogeneous, but live shots from the field are highly dynamic. Our video is parsed by means of a clustering technique, applied to both the spatial difference metric (SDM) and the histogram difference metric (HDM). The video for each news story file is parsed to locate the frame of studio-to-field transition. The frame number enables us to segment the audio track into an initial portion of anchor speech and subsequent portion of field speech. Results indicate that retrieval based only on the indexed anchor speech segments is as good as that based on the entire audio tracks.

In the future, we plan to improve our video parsing technique for story segmentation, and to collect a greater amount of data for experimentation.

ACKNOWLEDGMENTS

We would like to thank the Television Broadcasts Limited for providing the Cantonese news video in this project. We are grateful to Wai-Kit Lo from the Department of Electronic Engineering, for his help with audio indexing. This project is supported by a grant from the Hong Kong University Grants Council for the Area of Excellence in Information Technology.

REFERENCES

1. Bezdek, J. C., *Pattern Recognition with Fuzzy Objective Function Algorithm*, New York, Plenum, 1981.
2. Chien, L. F. and Wang, H. M., “Exploration of Spoken Access for Chinese Text and Speech Information Retrieval”, Proceedings of the International Symposium on Signal Processing and Intelligent Systems, 1999.
3. 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, 2000.
4. Lo, W. K., T. Lee and P. C. Ching, “Development of Cantonese Spoken Language Corpora for Speech Applications”, Proceedings of ISCSLP, Singapore, 1998.
5. Meng, H., W. K. Lo, Y. C. Li and P. C. Ching, “Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval”, Proceedings of ICSLP, 2000.
6. Salton, G. and M. McGill, M., “Introduction to Modern Information Retrieval”, McGraw-Hill, New York, 1983.
7. Singhal, A., and F. Pereira, “Document Expansion for Speech Retrieval”, Proceedings of SIGIR, 1999.
8. Wactlar, H., T. Kanade, M. Smith and S. Stevens, “Intelligent Access to Digital Video: Informedia Project”, IEEE Computer, Theme issue on Digital Library Initative, May 1996.
9. Wang, H. M., “Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching”, Proceedings of IRAL, 1999.