Multimedia Fusion in Automatic Extraction of Studio Speech Segments for 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 describes our progress in Cantonese spoken document retrieval. Over 60 hours of Cantonese television news broadcasts have been collected as part of AoE-IT Multimedia Repository. We have also developed the Multimedia Markup Language (MmML) for annotating the multimedia content in terms of anchor/field video frames and audio recordings. The audio tracks are indexed by a Cantonese syllable recognizer. Our investigation indicates that there is a large discrepancy in recognition performance, i.e. dropping from 59% to 39% in syllable accuracy (and corresponding reliability in audio indexing), as we move from anchor speech recorded in the studio to reporter/interview speech recorded in the field. Hence we present several automatic methods to extract anchor/studio speech from the audio tracks for retrieval: (i) extraction based only on video information using a fuzzy c-means algorithm; (ii) extraction based only on audio information using Gaussian Mixture Models; and (iii) a fusion strategy that combines video- and audio-based extraction. This paper presents the performance of various extraction techniques and the related retrieval performance in a known-item spoken document retrieval task.

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. We have been working on Cantonese spoken document retrieval based on local television news broadcasts [1]. Cantonese is a major dialect of the Chinese language, predominant in Hong Kong, Macau, South China and many overseas Chinese communities. This work presents a multimedia corpus that includes Chinese text, Cantonese audio and video derived from local television news broadcast. The corpus is annotated in terms of the Multimedia Markup Language (MmML). 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 to effect extraction: (i) video-based segmentation; (ii) audio-based segmentation and (iii) fusion of video- and audio-based segmentations. Previous work in this area include Mandarin (the major dialect of Chinese) spoken document retrieval by [2] and [3]; and the CMU Informedia project [4] which uses image and audio information concurrently for digital video access.

2. THE AOE-IT MULTIMEDIA REPOSITORY

The AoE-IT (Area of Excellence in Information Technology) Multimedia Repository is a collection of multilingual multimedia content that includes text, audio and video aimed to support relevant projects in the regional research community. The video corpus used in this work forms part of the AoE-IT Multimedia Repository and is collected from the Cantonese news broadcast from the Hong Kong Television Broadcasts Ltd. (TVB). Details of the video corpus used in this work are provided in Table 1. Each video file in the corpus contains a news story and comes with a text file that contains a textual summary with a title. The textual summary is by no means a verbatim transcription of the audio track. 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 (each Chinese character is pronounced as a syllable). The average title is 17.5 characters in length. An illustration is shown in Table 2.

Language	Cantonese Chinese
Source	TVB Jade channel
Digital Video Format	MPEG-1
Number of Stories	1627 (~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)
Minimum Length of News	4.5 sec
Maximum Length of News	8 min 55.0 sec

Table 1. The Cantonese video corpus, part of the AoE-IT Multimedia Repository. There are 1627 news stories in total.

<u>預科生可更改報讀學科優先次序</u>
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).

2.1 Annotation – Multimedia Markup Language (MmML)

We have designed the Multimedia Markup Language (MmML) for annotating content in the AoE-IT Multimedia Repository. MmML based on Synchronized Multimedia Integration Language (SMIL) 2.0 specifications. We have followed the XML schema hierarchy to design MmML as shown in Figure 1. There are three modules in the first level of the frame; they are Video, Audio and Text. We have adopted some elements (also known as tags) from SMIL's BasicMedia module MediaClipping module in our first-level modules. For instance, given that a video file contains video and audio tracks, our video module contains the elements VideoTrack and AudioTrack. Attributes are attached to elements to provide further description details. For example, there are different kinds of shots in VideoTrack – we use the attribute Name with value AnchorShot to label anchor shots. In this work it is important to distinguish among anchor (i.e. studio) versus field shots as well as anchor versus reporter/ interviewee speech, as will be explained later.

The news stories in our corpus typically begin with a report from the anchor(s) in the studio and are optionally subsequented by a live report from the field. All news stories have been manually annotated. Annotations include the name and gender of the anchor and reporter, start and end times of various news segments, temporal indices (in seconds) of changes in acoustic conditions, the speaking style and language/dialect of speech segments, etc.

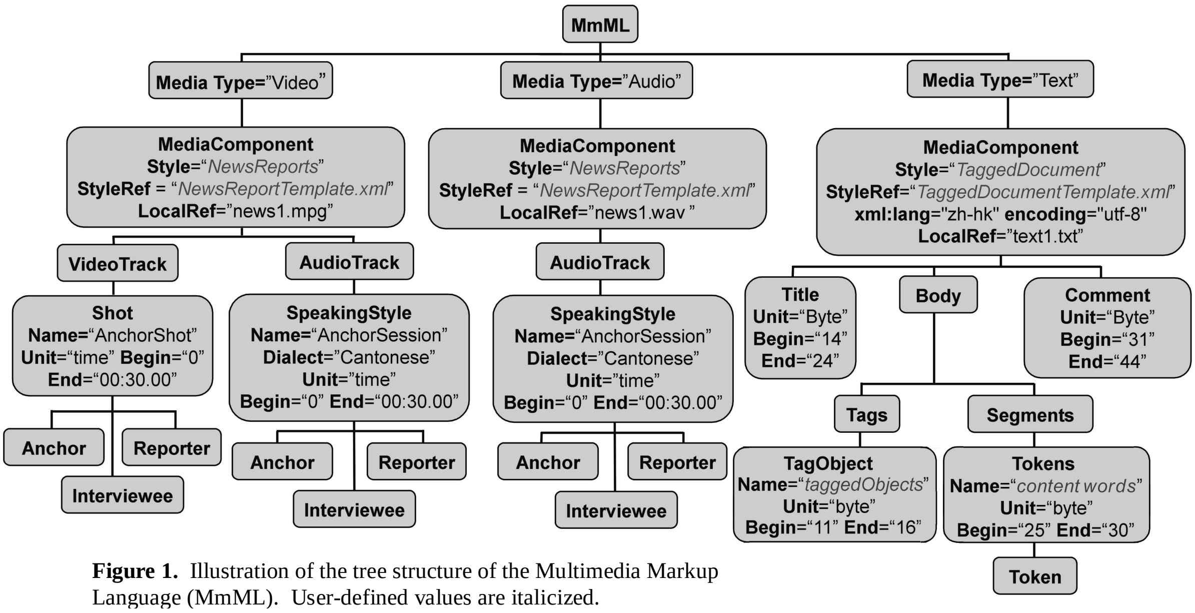
MmML
Media Type="Video"
Media Type="Audio"
Media Type="Text"
MediaComponent Style="NewsReports" StyleRef = "NewsReportTemplate.xml" LocalRef="news1.mpg"
MediaComponent Style="NewsReports" StyleRef = "NewsReportTemplate.xml" LocalRef="news1.wav "
MediaComponent Style="TaggedDocument" StyleRef="TaggedDocumentTemplate.xml" xml:lang="zh-hk" encoding="utf-8" LocalRef="text1.txt"
VideoTrack
AudioTrack
AudioTrack
Shot Name="AnchorShot" Unit="time" Begin="0" End="00:30.00"
SpeakingStyle Name="AnchorSession" Dialect="Cantonese" Unit="time" Begin="0" End="00:30.00"
SpeakingStyle Name="AnchorSession" Dialect="Cantonese" Unit="time" Begin="0" End="00:30.00"
Title Unit="Byte" Begin="14" End="24"
Body
Comment Unit="Byte" Begin="31" End="44"
Tags
Segments
Anchor
Reporter
Interviewee
Anchor
Reporter
Interviewee
Anchor
Reporter
Interviewee
TagObject Name="taggedObjects" Unit="byte" Begin="11" End="16"
Tokens Name="content words" Unit="byte" Begin="25" End="30"
Token

Figure 1. Illustration of the tree structure of the Multimedia Markup Language (MmML). User-defined values are italicized.

The MmML markups for our video corpus are automatically generated by a Java program that accepts an EXCEL file and an MmML template as input. The EXCEL file stores the manual annotations. The MmML template centralizes information for media components and can be extended by the use with new data types. Table 3 shows an example of an MmML video template file that includes the media components of video track. Table 4 shows an MmML markup of an MPEG-1 video file (including the video and audio tracks) together with its corresponding textual summary.

3. AUDIO INDEXING BY AUTOMATIC SPEECH RECOGNITION

The audio tracks of the video materials are indexed by automatic speech recognition. Audio tracks from the MPEG-1 video files of the news stories are extracted and converted to 16kHz mono-aural format. We have also developed a base syllable recognizer (no tone information in included) for indexing the audio tracks. Our recognizer is HMM-based and uses syllable initial (3-state HMMs) and syllable final (5-state HMMs) models. These models are right content-dependent HMMs with 16 Gaussians mixtures. The acoustic features used are 12 MFCC with the log energy and augmented with the first and second derivatives (39 parameters per vector). Details can be found in [1] and [5]. Thus far we have not attempted to improve audio indexing by developing gender-specific or speaker-dependent models, though the task is possible given the MmML annotations described earlier.

We have also transcribed about 2.75 hours of our audio tracks for the purpose of evaluation. The syllable accuracy of our recognizer is found to be 44.4%. The low accuracy is mainly due to harsh acoustic conditions (especially for audio recordings from the field) and the diverse speaking styles (read speech for the anchor versus spontaneous speech for the reporter/interviewee). To gauge the performance differences across various speaking styles and ambient conditions, we manually segmented and transcribed 20 audio stories (a subset of the 2.75 hours mentioned above) into anchor, and reporter/interview (i.e. field) speech. Syllable accuracies are shown in Table 5. We observe severe degradation in recognition performance as we move from anchor speech recorded in the studio towards reporter/interviewee speech recorded in the field. Audio indexing by recognition is computationally intensive and recognition performance affects spoken document retrieval performance. This motivates us to devise automatic methods for locating studio quality anchor speech from the audio tracks. In this way we aim to reduce the audio indexing effort required (by a factor of 4 in terms of the total duration of audio tracks), increase the recognition performance and audio indexing reliability, which should hopefully lead to better retrieval performance.

```
<MMML>
    <Template Type−"Video" Name−"NewsReports">
        <UserDefinedComponent Belongs−"VideoTrack" Type−"Shot"
        Name−"AnchorShot" MinOccurs−"1" MaxOccurs−"unbounded">
            <Item Name−"Anchor" Type−"person" />
        </UserDefinedComponent>
        <UserDefinedComponent Belongs−"VideoTrack" Type−"Shot"
        Name−"FieldShot" MinOccurs−"0" MaxOccurs−"unbounded">
            <Item Name−"Reporter" Type−"person" />
            <Item Name−"Interviewee" Type−"person" />
        </UserDefinedComponent>
    </Template>
</MMML>
```

Table 3. An illustration of an MmML template for video with the defined media component VideoTrack. Template Type in the schema is corresponding to the Media Type in MmML. User can define their own component by adding a new UserDefinedComponent and since we have two types of shot in VideoTrack, we have two UserDefinedComponent to describe them. MinOccurs indicates whether the component must be contained in media file or not, 1 means yes and 0 means not. MaxOccurs indicates the maximum number of a component being contained in a media file; unbounded means there is no limitation on this value.

```
<MmML>
<Media Type−"Video" Index−"1999080106">
    <MediaComponent LocalRef−"1999080106.mpg" Style−"NewsReports"
    StyleRef−"newsreports.xml" encoding−"big5" xml:lang−"zhtw">
        <AudioTrack Unit−"second" Begin−"0" End−"57">
            <SpeakingStyle Name−"AnchorSession" Dialect−"cantonese"
            Unit−"second" Begin−"0" End−"17">
                <Anchor Gender−"F">魏綺珊</Anchor>
            </SpeakingStyle>
            <SpeakingStyle Name−"ReporterSession" Dialect−"cantonese"
            Unit−"second" Begin−"17" End−"31">
                <Reporter Gender−"F">陳嘉怡</Reporter>
            </SpeakingStyle>
            <SpeakingStyle Name−"IntervieweeSession" Dialect−"cantonese"
            Unit−"second" Begin−"31" End−"57">
                <Interviewee Gender−"F" />
            </SpeakingStyle>
        </AudioTrack>
        <VideoTrack Unit−"second" Begin−"0" End−"00:57.00">
            <Shot Name−"AnchorShot" Unit−"second" Begin−"0" End−"17">
                <Anchor Gender−"F">魏綺珊</Anchor>
            </Shot Name>
            <Shot Name Name−"FieldShot" Unit−"second" Begin−"17" End−"57">
                <Reporter Gender−"F">陳嘉怡</Reporter>
                <Interviewee Gender−"F" />
            </Shot Name>
        </VideoTrack>
    </MediaComponent>
</Media>
<Media Type−"Text" Index−"1999080106">
    <MediaComponent LocalRef−"1999080106.txt" Style−"Essay"
    StyleRef−"abstract.xml" encoding−"big5" xml:lang−"zhtw">
        <Title Unit−"byte" Begin−"0" End−"18">菲律賓渡輪發生火警</Title>
        <Body>
            <Segments Name−"summary" Unit−"byte" Begin−"19" End−"86">菲律賓中
            部一艘渡輪發生火警，目前知道至少有三人死亡，百多人仍然失蹤。</Segments>
        </Body>
      </MediaComponent>
</Media>
</MmML>
```

Table 4. An illustration of an MmML markup for a news story with filing index 1999080106 (corresponding to the sixth story on August 1, 1999). The first layer shows a video file with its corresponding text file. The video file contains both tracks while the text file contains the textual summary of the news story.

Anchor	Reporter/Interviewee
59.3%	39.2%

Table 5. Syllable accuracies of audio indexing by base syllable recognition. Anchor speech is clearly articulated and recorded in the studio with favorable ambient conditions. Reporter/interviewee speech is spontaneous and recorded from the field, possibly with harsh acoustic conditions.

4. AUTOMATIC EXTRACTION OF ANCHOR/STUDIO SPEECH SEGMENTS

We have devised three automatic methods for extraction of anchor/studio speech segments. The first extraction 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 histograms. 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 right with an icon in the left; (iii) anchor on the left with icon on the right as well as (iv) two anchors side by side. 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].

This video segmentation algorithm is applied to the 1627 new stories in our corpus and evaluated against a hand-labeled reference. The annotators marked studio-to-field transitions based on changes in the video frames. The manual annotations indicate that 1545 of the news stories (~95.0%) contains a single studio-to-field transition and 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 with two seconds (i.e. 50 frames) from 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 6).

4.2. Audio-based Segmentation

We attempt to extract anchor/studio segments based on the audio information as well. The goal is to capture differences in the acoustic signal since studio speech tends to be less noisy when compared with field speech that may contain music, environmental noise, etc. We use single-state Gaussian Mixture Models (GMM) [7, 8] for audio-based segmentation. We trained one GMM to be the *studio model* and another to be the *field model* using the Baum-Welsh algorithm and 5 hours of audio data from our video corpus. The number of Gaussian mixtures was increased exponentially from 1 to 64. At 64 mixtures the GMMs can correctly extract most of the anchor/studio speech segments from the 5 hours of training data.

During testing, our audio-based extraction framework needs to distinguish news stories with no field shots from those with studio-to-field transitions. Hence for a given news story with T speech frames, we first compute the cumulative score by traversing with the studio model only:

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

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

$$Score_{studio\_to\_field} = \prod_{t=1}^{T_t}\sum_{i=1}^{64} w_i N_{studio}(\mu_i, \sigma_i) \prod_{t=T_t}^{T}\sum_{j=1}^{64} w_j N_{field}(\mu_j, \sigma_j)$$

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 consist 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 as in video-based segmentation. Results are shown in Table 6 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%) 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 in the boundaries labeled by audio-based segmentation in comparison with the video-based algorithm. We studied specifically 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.

	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 6. Performance of Video Parsing and Speech Classification.

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 improve automatic extraction of anchor/studio speech segments. Table 7 shows the *a priori* statistics with regards 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 7. Statistics (number of stories) from our AoE-IT video corpus in relation to the presence/absence of studio-to-field transitions in the audio and 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 detect a studio-to-field transition – we extract the anchor/studio segment according to the video-based algorithm, since its boundaries have deviate less from the reference boundaries (see Table 6).

Case 2: *Only* the video-based segmentation detects a studio-to-field transition – we extract the anchor/studio segment according to the audio-based algorithm since there exists news stories in this category (see Table 7).

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 7).

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 A Known-Item Retrieval task

We have formulated a *known-item retrieval task* based on our AoE-IT 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. Retrieval is based on the vector-space model in SMART [9] and details were described in [1]. Since there is only one relevant textual/audio document for each query, we adopt the average inverse rank (AIR) as our evaluation criterion:

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

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.

Each query / document is represented as a vector of syllable bigrams and skipped bigrams. This representation has previously been shown to give the best retrieval performance [1]. Figure 2 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 converted into syllable bigrams/skipped 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 2. Procedure for forming text-converted syllable bigrams / skipped bigrams.

5.2 Experimental Results

Retrieval based on text-converted syllable bigrams and skipped bigrams (i.e. from the textual summaries) provide 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 8 shows the retrieval results for various methods of extracting the anchor/studio speech segments. 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.

Extraction method for anchor/studio speech	AIR
Manual labeling based on video frames (reference)	0.633
Automatic video-based segmentation	0.628
Automatic audio-based segmentation	0.631
Fusion of video- and audio-based segmentation	0.641

Table 8. Spoken document retrieval performance based on extracted anchor/studio speech segments. The manual extraction result is included as a reference. Fusion of video- and audio-based segmentation gives the best retrieval results.

6. CONCLUSIONS AND FUTURE WORK

This paper reports on our progress in Cantonese spoken document retrieval. We describe a video corpus of over 1600 news stories from television broadcasts that we have collected as part of the AoE-IT Multimedia Repository. We aim to make the Repository available to support regional multimedia research projects. We have also developed the Multimedia Markup Language (MmML) for annotating the multimedia content to support this work and future related work. MmML is XML-based and is extensible to include a variety of annotations for multimedia content. We have manually annotated our video corpus in terms of anchor and field shots, anchor and field audio recordings, identity and gender of the anchors, language and dialect, etc. Audio segments need to be indexed for retrieval. We have developed a Cantonese syllable recognizer for indexing the audio tracks in our video corpus. Our investigation indicates that there is a large discrepancy in recognition performance, i.e. dropping from 59% to 39% in syllable accuracy (and corresponding reliability in audio indexing), as we move from anchor speech recorded in the studio to reporter/interview speech recognized in the field. Hence we developed several automatic methods to extract anchor/studio speech from the audio tracks for retrieval: (i) video-based segmentation aims to distinguish between the more homogeneous studio shots from the more dynamic field shots; (ii) audio-based segmentation uses a Gaussian Mixture Models (GMM) to distinguish the cleaner studio recordings from the more noisy field recordings; and (iii) a fusion strategy that combines video- and audio-based segmentation to achieve better extraction of anchor/studio speech. Fusion gave the best spoken document retrieval performance, given AIR=0.683. Future investigation will be devoted to the use of (noisy) field speech to further improve retrieval performance.

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. We thank Man Cheuk HO and Tin Hang LO for their assistance.

REFERENCES

1. 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.
2. 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.
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. Wang, H. M., "Retrieval of Mandarin Spoken Documents Based on Syllable Lattice Matching," *Proceedings of IRAL*, Taipei, Taiwan, 1999.
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. Reynolds, D., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, Elsevier Science, 1995.
8. 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.
9. Salton, G. and M. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, New York, 1983.