Robust Methods for Chinese Spoken Document Retrieval

HUI Pui Yu

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Systems Engineering and Engineering Management

Supervised by

Professor Helen M. Meng

August 2003


香港中文大學
圖書館系統
30 APR
UNIVERSITY
LIBRARY SYSTEM
THE CHINESE UNIVERSITY OF HONG KONG

Abstract of thesis entitled:

Robust Methods for Chinese Spoken Document Retrieval

Submitted by HUI Pui Yu

for the degree of Master of Philosophy

at The Chinese University of Hong Kong in August 2003

This thesis focus on the development of robust methods for Chinese spoken document retrieval (SDR). SDR is a technique that enables retrieval of relevant information from archives of spoken data. Robust methods refer to the methods that are tolerant of different disturbing factors and maintain (or even enhance) the retrieval performance. One of the disturbing factors to SDR is quality of the speech data. The spoken documents in our news archive are indexed by automatic speech recognition (ASR) of Cantonese base syllables. The news data usually begins with a report by the anchor(s) in the studio, followed by a live report by reporter(s)/interviewee(s) from the field. Recognition performance degrades significantly as we migrate from studio-quality speech (anchor speech) to field speech (reporter/interviewee speech). Recognition errors affect retrieval performance. Hence our investigation focuses on the fusion of audio and/or video information for the extraction of anchor speech and use recognition hypotheses to enrich the document representations. We formulated a known-item retrieval task and the experiments are performed using vector-space model (VSM). Evaluation is based on average inverse rank. Two robust techniques are investigated to improve the retrieval performance: (i) extraction of anchor speech using audio and video information and (ii) docu-

ment expansion using N-best recognition hypotheses and selected field speech segments. Using these robust methods can reduce the required indexing effort and improve the retrieval performances by 10%.

The third robust method we have explored is query expansion (QE). QE is a process aimed at reducing the query/document mismatch by expanding the query using words or phrases with similar meaning. The QE algorithm usually use to retrieve relevant documents from another collection of newswire text so as to avoid the inclusion of recognition errors. Previous work demonstrated that QE is beneficial to monolingual SDR. Therefore, we extend our work to cross-language SDR (CLSDR). CLSDR is a technique that enables retrieval of relevant documents in one language using queries in a different language. In this work, the CLSDR task retrieves Mandarin broadcast news data using English textual news data. We applied QE based on pseudo relevance feedback (PRF) to the CLSDR task. In PRF, the relevant terms from the initial retrieval output are used to expand the query for the second retrieval iteration. Retrieval is also based on VSM and is evaluated using mean average precision. Results show that PRF improves the retrieval performance of CLSDR by 25%.

摘要

這篇論文集中研究一些有助中文語音文件檢索 (spoken document retrieval) 的穩固方法。語音文件檢索是一項從語音資料庫搜尋相關資料的技術。“穩固”的意思是指檢索的方法是能夠容忍一些影響檢索性能的因素，包括語音文件的質素。新聞資料庫內的語音文件曾經用自動語音識別技術處理，並以廣東話基本音節 (base syllable) 作索引。新聞故事大抵是以演播室內的報導 (anchor speech) 開始，並以演播室外的報導 (reporter/interviewee speech) 作結。當我們比較語音識別技術的表現時，發現此技術在演播室外的表現較在演播室內的遜色。因語音識別技術的表現會直接影響檢索系統的性能，所以我們集中研究音像及/或錄影資溶合技術，從而抽取出演播室內的語音資訊。我們亦運用語音識別假設 (recognition hypotheses) 去增加文件的代表性。我們設定了一個已知項的檢索任務 (known-item retrieval task)，在音節空間上以向量空間模型 (vector space model) 進行檢索任務，並以平均倒轉等級 (AIR) 作為評估量度。我們探討了兩個方向的穩固方法：(一)以音像及錄影資訊去抽取演播室內的語音資訊；(二)以語音識別假設及演播室外的部份語音資訊作文件擴展。這些方法不單能減少所需要的標籤工作量，更能把檢索性能提升了10%。

第三個穩固方法的探討方向則是查詢問句擴展 (query expansion)。查詢問句擴展以近義詞或近義詞組去擴展查詢，目的是減少查詢及文件之間的不相配。為避免包括語音識別錯誤在查詢中，查詢問句擴展通常是用檢索另一文本資料庫的結果作擴展。從前的研究結果發現，查詢問句擴展對單語言語音文件檢索是有利的。因此，我們延伸了單語言音文件檢索的研究到跨語言語音文件檢索 (cross-language spoken document retrieval)。跨語言語音文件檢索是指使用者能以某一語言的查詢去搜索另一語言的語音文件。我們嘗試以英文文字查詢檢索普通話語音文件，並用準相關回饋 (pseudo relevance feedback) 作查詢問句擴展：利用原查詢問句檢索出之一組文件，不經使用者判斷即假定某一數量的文件皆為

相關，而這些假定的相關文件即經由相關回饋的程序重新建構查詢問句，再利用已重新建構的詢問做進一步的檢索。檢索亦是以向量空間模型 (vector space model) 進行，並以平均倒轉等級 (AIR) 作為評估量度。結果顯示準相關回饋能把跨語言語音文件檢索的性能提升25%。

Acknowledgements

I would like to thank my supervisor, Professor Helen Meng, for her guidance throughout this research project. Her comments and feedback are invaluable to my research and thesis writing. I also thank Helen for her experience sharing, all the training and opportunties she gave me. She talked with me when I had emotional upset, discussed with me when there was a puzzle to me and gave me chances to learn from the experts in this research area. She puts trust in us, gives us freedom to choose our research topics and ways to achieve them. She also provides us with excellent computing resources and facilities so that we can focus on our research and schoolwork.

I also want to thank my entire thesis committee, Professor Helen Meng, Professor Wai Lam, Professor Christopher Yang and Professor Kui-Lam Kwok, for their time, effort and valuable advice. I would like to express my gratitude to Professor Pak-Chung Ching for his suggestions and Professor Sean Tang for his teaching in this work. Besides, I had the good fortune to gain knowledge from Dr. Wai-Kit Lo, Dr. Hsin-Min Wang and Mr. Yuk-Chi Li, who spent much time and patience to teach me and provided me a lot of helpful suggestions.

Members in Human-Computer Communications Laboratory helped me a lot too. I want to thank Dias and Florence for being my team mates in the final year project; Homa and Michael Lo for their help in the development of MmML; Kon-Fan, Kin and Tony for providing me some useful programs and scripts; Ben, Bonnie, Kui and Michael Lau for their share of TA workload; Ada,

Brenda, Julia, Silvia and Tiffany for their advices on academic and research issues; Simon for his encouragements, reminders (and comments), and the interesting ideas raised; Chat, Edmond, Ka-Fai, May, Sunny and Winnie for their time in chitchat.

Friends, including Agatha, Amelia, Anthony, Brian, Carol, Edward, Eliza, Ka-Chun, Kenneth, Mr. and Mrs. Yeung, support and bring happiness to me all the time. Friends from ASES Stanford formed valuable friendships with me and brought me an unforgetable experience. Students from the course SEG3510 (Human-Computer Interaction) and the helpers let me have good memory too. Thanks for all the joyful time they gave me.

Aggie, Arthur, Esther, Iris, Mandy, Monica, Mr. and Mrs. Leung and technical staff from our Department are very nice and I am appreciative of their support in these years.

Finally, I would like to thank KT for his advices, encouragements, reminders and sharing throughout these years; my brother Ray and his best friend Jess for their wordless support to all the decisions I have made; and most importantly, my Mum for her endless love and the confidence she continuously gives me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information retrieval (IR) tackles the problems of organization, representation, storage, accessibility and retrieval of information. Information in different disciplines has become available on the Internet and accessible for the general public, for example, supreme court multimedia [1], significant political and historical events and personalities [2]. Information sharing is a form of communication. Information available is not only limited to text and speech, but also music, video, image and graphics. Information can be accessed across the barriers of location, time and language: users can access information from other countries, in other languages and/or those happened in the past [3]. While there is a vast amount of multilingual and multimedia information available on the Internet, there is a potential need for suitable formalizations and in flexible / extensible representations. Much research work has been focused on the standardization of structure [4] and automation of the standardization process [5]. However, for research work in IR, there is a need of a structure / a set of markups that is able to index the information in categorization hierarchy.

IR technology enables the user to retrieve the documents that are relevant to the input and retrieve as precisely as possible [6]. Conventional IR systems are mostly focused on textual information. Aside from text, there is vast

amount of information in spoken form, e.g. from television and radio broadcasts [7]. Hence there is an increasing emphasis on retrieval with multimedia information and one of them is spoken document retrieval (SDR) [8].

Estimation showed that more than 80% of web pages are written in English [9] and much research effort has focused on English content. However, Chinese is predicted to be another predominant language to be used by the Internet population by 2005 [10]. Information available comes in different languages. Users not only want to retrieve information in their native languages, but also in other languagues available. The situation raises an interesting research issue in Chinese and cross-language retrieval.

1.1 Spoken Document Retrieval

SDR is a task that transcribe, index and retrieve a collection of spoken documents. The SDR system usually uses user-specified textual queries to retrieve spoken documents. Textual queries have been transcribed into syllables by pronunciation dictionary lookup. Spoken documents have been transcribed and indexed by means of automatic speech recognition (ASR). ASR is used to extract information from spoken documents and present them as transcriptions in syllables representation. A retrieval engine is responsible for the retrieval of relevant information based on the transcribed queries from the user. A list of relevant documents is returned to the user as retrieval output. An overview of a SDR system is shown in Figure 1.1. Details of the ASR and IR model will be presented in Chapter 3.

The objective of research in SDR is to retrieve spoken documents to meet user's requests. With the help of SDR, users can browse through spoken documents. The spoken documents may be telephone voicemail data, recordings of lectures, meetings, radio broadcasts / broadcast speech or audio tracks of video documents / video broadcasts.

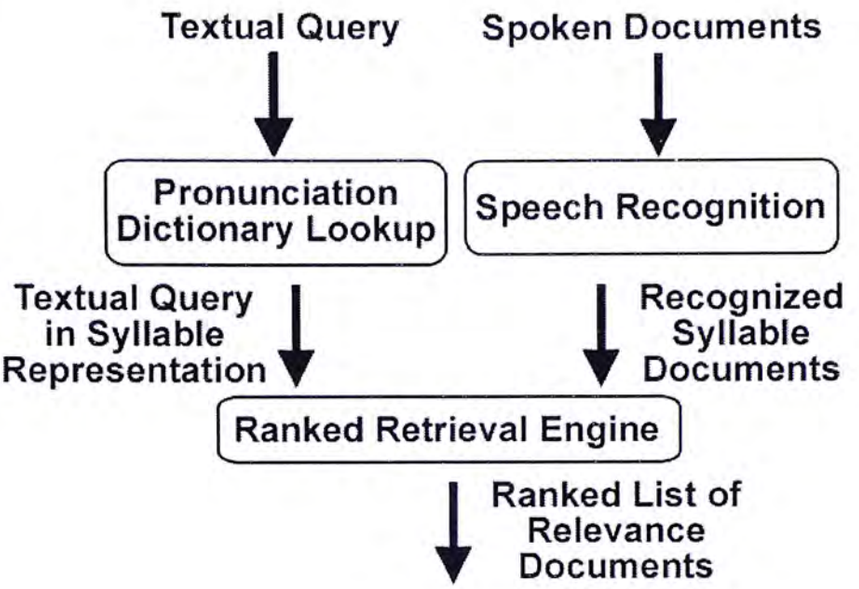

Figure 1.1: An overview of a SDR system, which illustrates the use of pronunciation dictionary lookup and ASR techniques.

The first SDR task was introduced in the Sixth Text REtrieval Conference (TREC 6) in 1997 [11]. There are a variety of indexing methods in English SDR, including word-based, syllable-based and phoneme-based. Research communities active in the area of SDR include the Text REtrieval Conference (TREC) [12] and Topic Detection and Tracking (TDT) [13] communities, where efforts exist not only for English, but also French, German, Italian and Chinese. SDR has gradually become a world problem and needs to be applicable to many different languages: some of the efforts in SDR have been focused on Dutch [15], German [16], Indian [17], Japanese [18] [19], Portuguese [20], etc. Additional efforts in Chinese speech retrieval exist for Mandarin [21] [22] and Cantonese [23]. Besides, HP SpeechBot^{TM}has introduced SDR to the general public [14]. Users can search across radio programs with SpeechBot^{TM}. SpeechBot^{TM}has currently indexed 17,274 hours of radio programs from 28 websites. The radio programs cover 11 topics and are updated daily. The Informedia^{TM}project at CMU [24] involves retrieval of multimedia informa-

tion from digital libraries [25] [26], which have broadcast data in English and Croatian [27]. The Rough'n'Ready^{TM}audio indexing system at BBN [28] segments the audio files into sections and enables searching of audio sections based on speaker, topic or concept [29].

Recent developments in SDR include cross-language SDR (CLSDR) and multilingual SDR (MLSDR). CLSDR means that users can retrieve spoken documents in one language using queries in another language. MLSDR is a system developed for two or more languages. Illustrations for CLSDR and MLSDR have been shown in Figures 1.2 and 1.3 respectively. In CLSDR system, queries in one language have been translated into another language for retrieval. The queries and retrieved documents are in different languages. In MLSDR system, the documents in one language have been translated to form another collection of documents. Queries are used to retrieve translated documents and both translated and untranslated documents will be returned to users as retrieval output. Examples for CLSDR include the use of English text queries to retrieve Chinese spoken documents [30] and the retrieval of German spoken documents in response to French text queries [31]. LODEM [32] and LIMSI-CNRS [33] are examples for MLSDR. LODEM is a multilingual lecture-on-demand system, which searches relevant segments of video across Japanese and English. There is a SDR system in LIMSI-CNRS, which is developed for American English, Arabic, French, German, Mandarin, Portuguese and Spanish.

A popular approach to SDR is to couple word transcription that uses large-vocabulary continuous speech recognition (LVCSR) with IR techniques [34]. However, word transcription inevitably encounters the open vocabulary / out-of-vocabulary (OOV) problem in recognition. The OOV problem refers to the existence of *unknown* words in the spoken audio. An *unknown* word refers to a word that is absent from the recognizer's vocabulary. The use of subword-

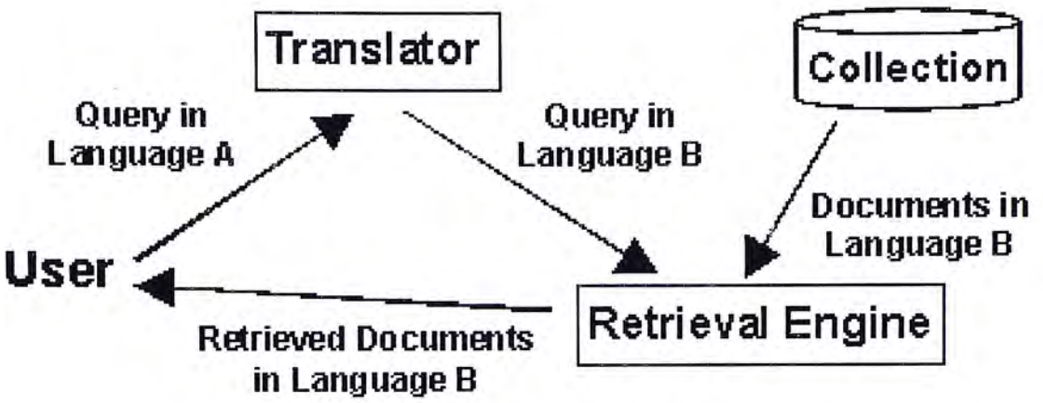

Figure 1.2: An overview of a CLSDR system. The query in one language has been translated into another language. The retrieved documents are in the language different from the query.

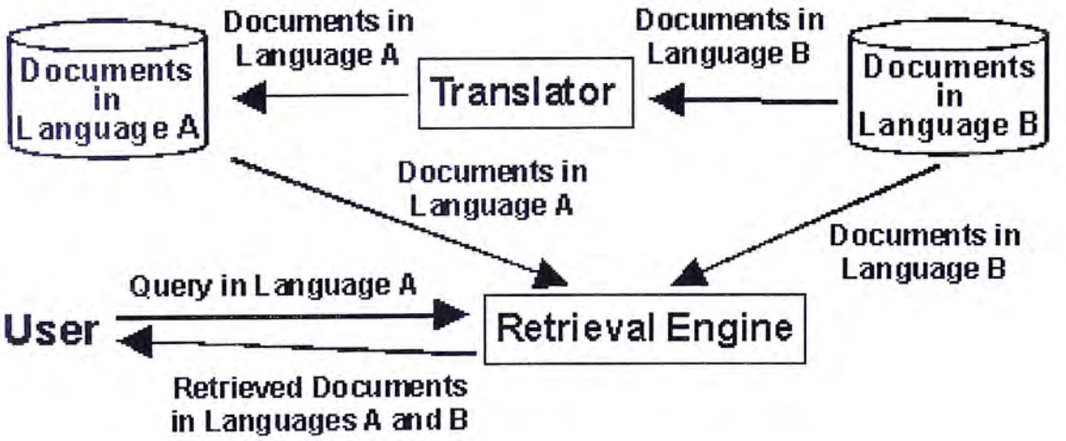

Figure 1.3: An overview of an MLSDR system. The documents have been translated and stored in the databases for retrieval. Both translated and original documents will be sent to the user as retrieval output.

based units (e.g. phoneme, syllable, characters, etc) can handle the OOV problem. The spoken documents may be indexed with phoneme n-grams [35] [36] [37] or phone lattices [38]. Subword-based indexing can enhance recall by providing complete phonological coverage of the spoken audio and thus overcome the OOV problem. Subword-based indexing with the syllable unit is particularly suitable for the Chinese language because the Chinese language is monosyllabic in nature. As we will explain in the Section 1.2, a finite set of syllables is sufficient for the indexing of Chinese audio documents. However, syllables do not contain any lexical information. Single base syllables may have the ambiguity in pronunciation because tone information is missing in base syllable. The lack of lexical information of syllables may harm the retrieval performance. Previous work in Chinese text retrieval [39] showed that retrieval base on syllable n-grams can have comparable result with word-based retrieval. The sequential constraints of syllable n-grams can partially compensate for the lack of lexical knowledge when compared to word-based retrieval.

In addition to SDR, other work in multimedia retrieval has been done on video image information. Scene breaks [40], blank frames [41], frame similarity [42] (computed from color histograms, spatial histograms or eigenface similarities), phrase templates [41] and exploitation of image information [43] have been used for automatic story segmentation.

1.2 The Chinese Language and Chinese Spoken Documents

As the Chinese language is becoming another predominant language for Internet population, this work on SDR aims to index and retrieve Chinese spoken documents. In response to the research interest in Chinese SDR, our approach takes into consideration the linguistic properties of the Chinese language. Chi-

nese has many dialects, each characterized by their differences in phonetics, vocabularies and syntax. However, all of the Chinese dialects are monosyllabic in nature. Each syllable carries a lexical tone. For example, Cantonese is one of the major Chinese dialects. There are about 600 base syllables in Cantonese and each base syllable has between 6 to 9 lexical tones. Together they form about 1,600 distinct tonal syllables that can fully characterize Cantonese phonology.

In its written form, Chinese is a sequence of characters. Each character is pronounced as a tonal syllable. "Tonal" means that tone information is included. A Chinese syllable can be decomposed into an *initial* and a *final*. *Initial* refers to the optional onset consonant in the first part of a syllable. *Final* consists of the vowel / diphthong (nucleus) followed by an optional coda consonant in the syllable. Figure 1.4 shows the typical syllable structure of a Chinese syllable [44].

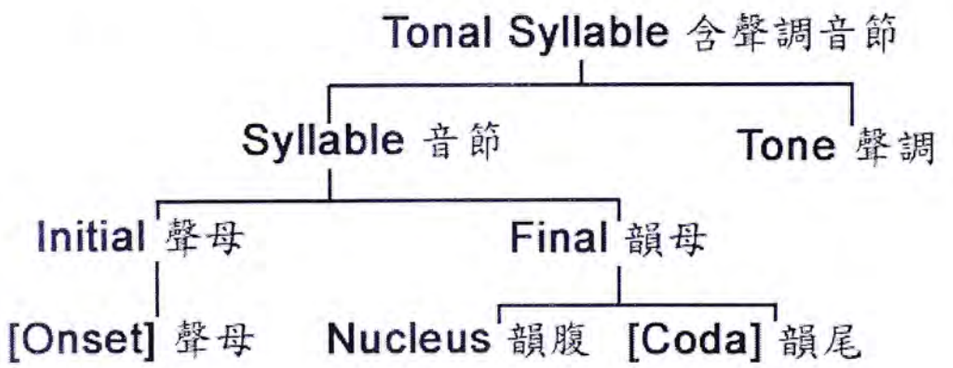

Figure 1.4: General structure of the Chinese syllable. The components in a pair of square brackets are optional consonants in a Chinese syllable.

The character-syllable mapping is many-to-many. On one hand, a given character may have multiple tonal syllable pronunciations – for example, the character 樂 may be pronounced as /ngaau6/[2], /ngok6/, /lok3/ or /lok6/ in Cantonese. A detailed example of multiple pronunciations is given in Ta-

[2]A Cantonese tonal syllable, the number encodes the tone of the syllable.

ble 1.1. On the other hand, a given tonal syllable may correspond to multiple characters. Consider the pronunciation /zi1/, which corresponds to 79 Chinese characters as shown in Table 1.2. A two-syllable pronunciation /baan1 zoeng2/ can refer to two different sets of Chinese words given in Table 1.3.

Chinese words	**Cantonese syllable**	**Meaning**
樂業	/**ngaau6** jip6/	respect your own occupation
音樂	/jam1 **ngok6**/	music
樂	/**lok3**/	a Chinese surname
快樂	/faai3 **lok6**/	happy

Table 1.1: An example shows the multiple Cantonese pronunciations of the character 樂 and their corresponding words, syllables and meanings.

Pronunciation	**Possible Chinese characters**
zi1	次 之 知 資 支 齊 氏 枝 姿 脂 滋 肢 茲 芝 吱 蜘 孜 諮 恣 齜 錙 髭 咨 吡 孳 輜 趑 胝 卮 玆 祇 梔 淄 貲 緇 甾 秖 胑 衹 偨 嵫 椥 粢 菑 搘 鄑 榰 鳷 鼒 鎡 鮨 鯔 齍 泜 仔 砥 孖 疻 胔 觶 訾 齎 疷 秶 栺 胾 澬 汥 崰 璾 椔 茬 紎 秪 粢 跂 鈭 鶅 衼

Table 1.2: Examples on single Chinese character homophones correspond to the pronunciation /zi1/.

The Chinese characters that have more than one pronunciations or meanings are called homographs. Homophones are the different Chinese characters that sound alike. In addition to homographs and homophones, another source of ambiguity in the Chinese language is the definition of a Chinese word. A word may consist of one or more characters. There are approximately 13,000 characters in Chinese according to the BIG-5 character set and possible ways of deriving new words from characters are legion. The problem of identifying the word string(s) in a character sequence is known as the *word segmentation* or

Chinese words	Meaning
班長	a class leader
頒獎	give award to

Table 1.3: Examples on homophones and their corresponding meanings of the two-syllable pronunciation /baan1 zoeng2/.

word tokenization problem. Consider the example in Table 1.4, the underlined word strings can be segmented in two different ways. Different segmentations produce different word sequences and therefore with different meanings. Segmentation 1 is the correct segmentation according to the meaning of the original sentence.

Chinese character string	第二屆立法會議席
Cantonese tonal syllable string	/dai6 ji6 gaai3 lap6 faat3 wui2 ji5 zik6/
Segmentation 1 (meaning of the underlined words)	第二屆 立法會 議席 ⇐ Correct! (number of official members in Legislative Council)
Segmentation 2 (meaning of the underlined words)	第二屆 立法 會議 席 ⇐ Incorrect! (legislative assemblies)
Overlapping character bigrams	第二 二屆 屆立 立法 法會 會議 議席
Overlapping syllable bigrams	/dai6_ji6/ /ji6_gaai3/ /gaai3_lap6/ /lap6_faat3/ /faat3_wui2/ /wui2_ji5/ /ji5_zik6/

Table 1.4: An example on the word tokenization ambiguity. Different segmentations carry different meanings. The word tokenization problem can be avoided when overlapping bigrams are used.

Consideration for the ambiguities led to a previous effort investigating the "optimal" indexing / retrieval unit for Cantonese spoken documents [23]. Base

syllables provide full phonological coverage of the Chinese audio. The use of base syllables (instead of tonal syllables) avoids tone recognition errors in Canontese speech recognition.[3] ASR uses base syllable can attain higher recognition speed. The most effective indexing / retrieval unit was found to be overlapping syllable bigrams [23]. The overlap can circumvent the problem of word tokenization ambiguity, as shown in Table 1.4. The use of n-grams provides sequential constraints that partially capture lexical information to some degree. Among the n-grams studied (i.e. unigrams, bigrams and trigrams), bigrams gave the best retrieval performance. The observation of retrieval with bigrams [23] agrees with results from other studies in text retrieval, e.g. for Mandarin Chinese [45] and for Chinese text [46]. Skipped syllable bigrams have also been used to capture Chinese abbreviations as proposed in [47]. In a four-character window, a skipped bigram is formed by taking the first character, skipping the second and taking the third. Another skipped bigram is formed by taking the second character, skipping the third and taking the fourth. For the syllable sequence (s1, s2, s3, s4), the skipped syllable bigrams are (s1s3, s2s4). Many Chinese abbreviations are derived from skipping characters, e.g. 香港中文大學 (The Chinese University of Hong Kong) can be abbreviated as 中大 (taking out the third character, skipping the fourth and taking the fifth). Moreover, synonyms often differ by one or two characters/syllables, e.g. both 中華文化 and 中國文化 mean "Chinese culture".[4] Hence skipped syllable bigrams can contribute towards retrieval performance.

[3] Investigation [23] reveals that if we had perfect tone recognition, the tonal information would lead to performance gains in retrieval. However, the net effect based on actual (imperfect) tone recognition is only marginal.

[4] This example is borrowed from [30].

1.3 Motivation

The multimedia corpus used in this work includes Chinese text, Cantonese audio and video. The need of an efficient management of the corpus motivated us to design a set of markups / a structure. It enables us to embed the description of the content of the multimedia data in a textual format.

Recognizers perform much better on speech recorded in studio than in the field. This is because the acoustic condition is relatively fixed and studio-quality speech usually involves speech from no more than two anchors. The acoustic conditions change a lot in the field speech and it involves many segments of non-speech sounds and interviews with different speaking styles. This motivates us to derive automatic methods to extract studio speech and selected field speech segments so as to minimize the adverse effect of speech recognition errors and information loss due to extraction respectively. The investigation on the automatic extraction of studio-quality speech is demonstrated using monolingual SDR with a focus on Cantonese.

Research in IR has suggested that speech recognition errors degrade SDR performance. There are research works focused on the development of robust retrieval approaches against recognition errors. The approaches include multiple recognition hypotheses [47] and acoustic confusion [48] to expand either document representations for robust matching of indexing terms. We attempt to work on document expansion using N-best recognition hypotheses. The approach suggested in this work aims to involve multiple recognition hypotheses to enrich document representation.

The importance of the indexing terms can be reflected by the weights apply on them. Different weighting schemes are proposed based on the occurrences of recognition hypotheses and recognition accuracies of the speech segments. The effect of the weighting schemes on retrieval performance has been investigated.

1.3.1 Assisting the User in Query Formation

Users are in lack of knowledge regarding the specific domain / content of information. For example, if the user wants to search for help information in Chinese, he can use the query 幫忙(meaning: help). However, in the computer domain, "help" is usually translated into 救助 or 說明 . The user takes time to go through a few iterations so as to formulate suitable queries for retrieval purposes. In order to save the user's time for query refinement, there is an increasing need of automatic query modification so as to optimize the user's query for retrieval in a specific domain / content. This process may help user to retrieve more relevant information and improve the retrieval performance. The work in [48] showed that query expansion can bring improvement to monolingual SDR. Therefore, this work extends the work in [48] from monolingual SDR to CLSDR. In CLSDR, we use English queries to retrieve Mandarin spoken documents. We have explored the use of pseudo relevance feedback (PRF) for query expansion. PRF is an automatic feedback method. The aim of PRF is to improve the retrieval performance by expanding the query with terms from the top-ranking documents. PRF is able to emphasize some relevant terms and de-emphasize some non-relevant terms in the original query.

1.4 Goals

The main goal of this work is to develop robust SDR techniques for both monolingual (with a focus on Cantonese) and cross-language (retrieve Mandarin news using English queries) retrieval tasks. Robust SDR techniques refer to the techniques that are tolerant of different disturbing factors and maintain (or even enhance) the retrieval performance. Examples of the disturbing factors include the quality of speech data and automatic speech recognition errors. This work focuses on investigating the fusion of different sources of

information for SDR. Information can come from the audio and video tracks of the broadcast news data, Cantonese base syllable recognizer and retrieved list of relevance documents. We shall apply different information fusion approaches to the SDR task and address the following research issues:

- How can we use the audio and video information from the Cantonese news archive to perform anchor speech extraction? How well do they perform? What are their effects on monolingual SDR performance?
- How can we use the information from the recognition output and field speech to enrich document representation? How can they affect the monolingual SDR performance?
- How can we use the retrieval output to help the user to refine his queries? What is the retrieval performance after applying the automatic query refinement technique to cross-language task?

1.5 Thesis Organization

This thesis is organized as follows: Chapter 2 provides detailed information about the experimental corpora, especially the AoE-IT Multimedia Repository and Multimedia Markup Language. Chapter 3 talks about the components in a SDR system and the baseline results of the retrieval performance. Chapter 4 describes the methods of the automatic anchor/studio speech extraction, their performances and the effects of different segmentation methods on our speech retrieval results. The techniques of video parsing, speech classification, fusion of video- and audio-based segmentation information and our fusion strategy will be included in Chapter 4. Chapter 5 presents our document expansion techniques and their influences on our retrieval results. Document expansion techniques include the use of selected field speech segments and N-best recognition hypotheses. Chapter 6 talks about our work on CLSDR with an

emphasis on query expansion using PRF. Finally, Chapter 7 concludes and discusses possible future directions.

Chapter 2

Multimedia Repository

This chapter presents the detailed information about the experimental corpus used in this work. The multimedia corpus we have collected is AoE-IT Multimedia Repository. AoE-IT Multimedia Repository is a collection of multimedia Internet content that has been organized, transcribed and annotated to support research in various information technologies. The Multimedia Markup Language (MmML) is a convention for annotating multimedia in order to support research efforts in bilingual text retrieval as well as video / audio retrieval. The design of MmML will also be presented in this chapter.

2.1 The Cantonese Corpus

The AoE-IT (Area of Excellence in Information Technology) Multimedia Repository [49] is a collection of multilingual multimedia content, which includes text, audio and video. The Cantonese news data in the repository are derived from the Cantonese news broadcasts from the Jade [50] channel of the Hong Kong Television Broadcasts Limited (TVB). The archive includes news stories in the *RealMedia*^{TM}[51] and the *MPEG-1* formats [52]. The news stories are manually segmented from television news programs and the temporal structure of a typical news program is illustrated in Figure 2.1. Each news story

is a single video clip accompanied by a brief textual summary with a story title. Figure 2.2 shows an example of the textual summary of a news story, together with its title (underlined). However, the summary is not a verbatim transcription of the audio track of the news story. We have found that the length of the textual summary is roughly a quarter that of the audio track, measured in the number of characters / syllables. The average length of the summary titles is approximately 16.5 characters.

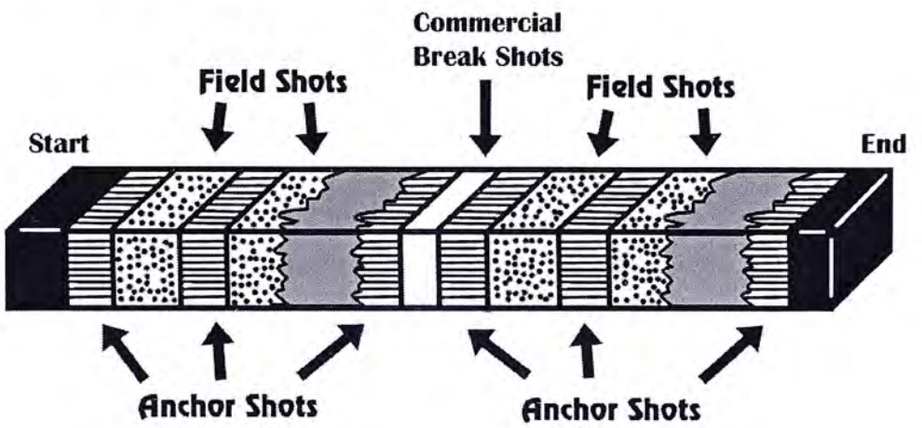

Figure 2.1: The temporal structure of a television news program.

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。

Figure 2.2: An example of the textual summary of a news story together with its title, which is underlined, from our corpus.

Very often, the news story begins with a report by the anchor(s) in the studio, followed by a live report by reporter(s)/interviewee(s) from the field. The anchor shots are relatively homogeneous – there are only four patterns of anchor shots can be found in our video archive, as shown in Figure 2.3.

However, there is no fixed pattern for field shots. The patterns found in anchor shots are key features for the single-link algorithm in the video processing, which will be report in detail in Chapter 4.

Figure 2.3: The four typical patterns of anchor shots in our entire video corpus.

2.1.1 The *RealMedia*™ Collection

The *RealMedia*™ data we have collected covers a six months period from July 1999 to December 1999. Details of the video corpus are provided in Table 2.1. The *RealMedia*™ data has a frame size of 160 × 112 pixels and a frame rate of 15 frames per second (fps). Table 2.2 shows the encoding details of the

RealMedia™corpus.

Language	Cantonese Chinese
Source	TVB Jade channel
Digital video format	*RealMedia*™
Number of stories	1,722 (around 39.9 hours)
Extraction period	7 July to 31 December 1999
Average length of news	1 min 23.43 sec (per story)
Minimum length of news	7.9 sec
Maximum length of news	8 min 13.8 sec

Table 2.1: Detailed information of the Cantonese video corpus in the *RealMedia*™format.

Frame size	160 × 112 pixels
Encoded frame rate	15 fps
Video codec	33.0 kbps (*RealVideo*™)
Sampling frequency	8kHz
Audio codec	12 kbps music (*RealAudio*™)

Table 2.2: Encoding information of the Cantonese video corpus in the *RealMedia*™format.

2.1.2 The *MPEG-1* Collection

Another batch of video archive we have collected is in the *MPEG-1* format. The corpus covers a period of four months. Table 2.3 shows the details of the video corpus. The *MPEG-1* data has a frame size of 352 × 288 pixels, which is four times larger than the *RealMedia*™data and a frame rate of 25

fps. Table 2.4 shows the details of the encoding of the *MPEG-1* video archive. Comparison between Table 2.2 and Table 2.4 suggests that the *MPEG-1* data is in higher quality and therefore all the experiments in monolingual spoken document retrieval (SDR) are based on the *MPEG-1* collection only.

Language	Cantonese Chinese
Source	TVB Jade channel
Digital video format	*MPEG-1*
Number of stories	1,627 (around 60.4 hours)
Extraction period	7 July to 17 August 1999 5 October to 31 December 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 2.3: Detailed information of the Cantonese video corpus in the *MPEG-1* format. There are 1,627 news stories in total.

Frame size	352 × 288 pixels
Encoded frame rate	25 fps
Video codec	1,150 kbps (*MPEG-1* video)
Sampling frequency	44.1kHz
Audio codec	192 kbps (*MPEG-1 Audio Layer II*)

Table 2.4: Encoding information of the Cantonese video corpus in the *MPEG-1* format.

2.2 The Multimedia Markup Language

In order to present and store the multimedia data in a categorization hierarchy, we have designed the Multimedia Markup Language (MmML). The use of MmML is to annotate content in the AoE-IT Multimedia Repository. Metadata can present the details / contents of non-textual objects in textual format.

The design of MmML is based on the Synchronized Multimedia Integration Language (SMIL™) 2.0 specifications [53]. We have borrowed the definition of Continuous Media from SMIL™[54] to be the definition of our multimedia data – continuous media reders to "*Audio files, video files or other media for which there is a measurable and well-understood duration.*" We have also followed the XML schema hierarchy to design MmML as shown in Figure 2.4. An illustration of the comparison between SMIL™and MmML hierarchies is also shown in Figure 2.5. As indicated in Figure 2.4, there are three modules in the first level of the frame; they are VIDEO, AUDIO and TEXT. We have adopted some elements and attributes from SMIL™'s BASICMEDIA module and 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 between anchor (i.e. studio) versus field shots as well as between anchor versus reporter/ interviewee speech, as will be explained in details in Chapter 4.

The news stories in our corpus typically begin with a report from the anchor(s) in the studio and there is optional subsequent live report from the field. All news stories have been manually annotated. Annotations include the name and gender of the anchors and reporters, start and end times of various news

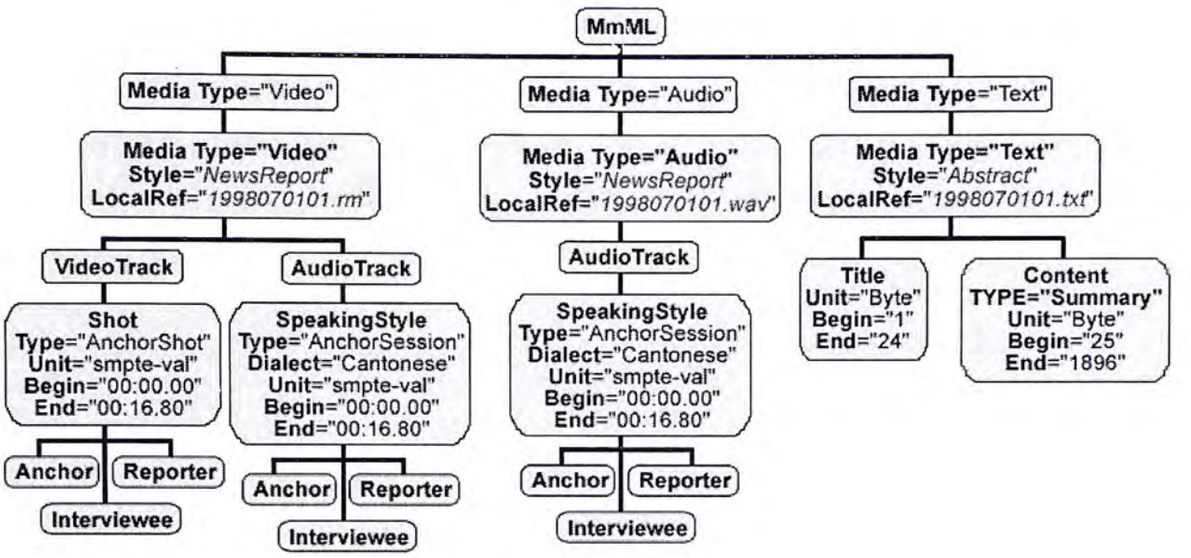

Figure 2.4: Illustration of the tree structure of the MmML. User-defined values are italicized.

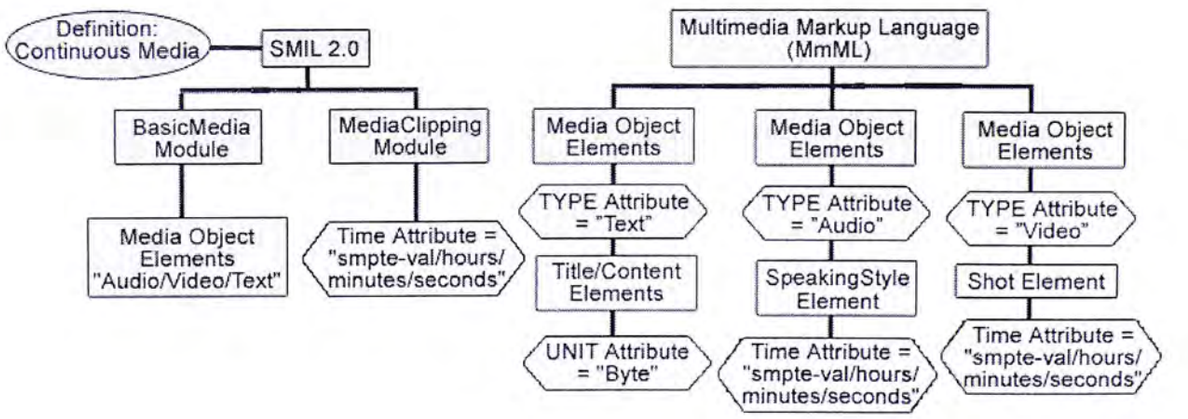

Figure 2.5: SMIL^{TM}2.0 Hierarchy versus MmML Hierarchy.

segments, temporal indices (in the format of smpte-val) of changes in acoustic conditions, the speaking styles and language/dialect of speech segments, etc.

The MmML markups for our video corpus are automatically generated by a Java program that accepts an EXCEL file and an associated XML schema as input. The EXCEL file stores the manual annotations. The file of XML schema centralizes information for media components and can be extended by the use with new element and attribute. Figure A.1 in Appendix A shows the XML schema of MmML. Figure B.1 in Appendix B illustrates a sample of MmML markup of an *MPEG-1* news file (including the video and audio tracks) together with its corresponding textual summary.

2.3 Chapter Summary

In this chapter, we present the details of the multimedia corpus to be used in our experiments, which is AoE-IT Multimedia Repository. The repository is a collection of multimedia and multilingual content, including audio, video and text. The Cantonese video news archive in the collection is provided by a local television station. In order to annotate the multimedia content in a meaningful and structural way, we have designed a convention of markup, namely Multimedia Markup Language (MmML). The design of MmML is based on SMIL™2.0 specification and followed the XML schema. With the MmML, we can present non-textual content / information in textual format.

Chapter 3

Monolingual Retrieval Task

This chapter presents the components of a spoken document retrieval (SDR) system and its baseline retrieval performance in this thesis. This includes the details of speech recognizer used for automatic transcription of spoken documents and the information retrieval model adopted in our work. An overview of a SDR system is shown in Figure 1.1.

3.1 Properties of Cantonese Video Archive

Recall that a news story often begins with a report by the anchor(s), followed by a live report by reporter(s) / interviewee(s) (see Figure 2.1). The anchor reports are primarily studio-quality audio delivered in Cantonese. Live reports consist mainly of spontaneous speech (e.g. from interviews). They may involve code switching among Cantonese, Mandarin and English (the languages of Hong Kong's trilingual environment). Moreover, the live reports are often recorded from highly variable acoustic conditions, e.g. with the reporter's voice-over, singing, music, applause, severe ambient noises, etc. Hence live reports generally present acoustic conditions that may be too harsh for reliable automatic speech recognition (ASR). Basically, we can classify the news stories into three main categories: (i) news stories with anchor-to-field transitions in

both audio and video tracks, (ii) news stories with anchor-to-field transition in video track only and (iii) news stories with anchor shots / audio only. An illustration of the three categories is shown in Figure 3.1. We have manually labeled each news story with a segment boundary that indicates a studio-to-field transition. Annotation is based *only* on the video frames.

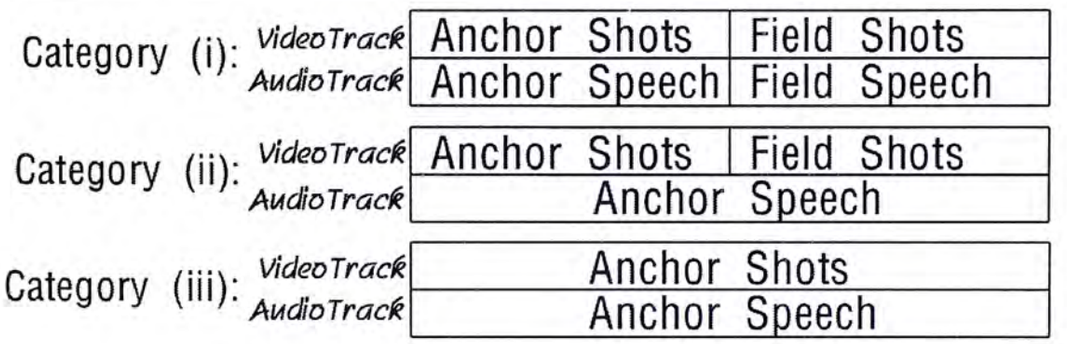

Figure 3.1: An illustration of the three categories of our news archive. They are (i) anchor-to-field transitions in both video and audio tracks, (ii) transitions in video track only and (iii) no transition from anchor to field in both tracks.

3.2 Automatic Speech Transcription

In SDR task, speech data has to be transcribed / indexed for further processing. This can be done by means of speech recognition technology. Speech transcription / indexing usually involves large-vocabulary continuous speech recognition (LVCSR) in the form of word units, for example, in transcribing of English spoken documents. However, as mentioned in Section 1.2, due to the linguistic properties of the Chinese language, we use base syllables as our indexing / retrieval unit for Cantonese spoken documents. In other words, we have designed a Cantonese base syllable recognizer to index the Cantonese audio.

3.2.1 Transcription of Cantonese Spoken Documents

Transcriptions for the Cantonese spoken documents (audio tracks of the video materials) used in this work are obtained by using a Cantonese base syllable recognizer developed ourselves [55]. The seed models of the recognizer is trained with 20 hours of data from CUSENT^{TM} corpus [56]. This corpus contains clean, phonetically-rich, continuous read speech recorded in a sound-proof room with a microphone. Since the training data have very different properties with our experimental corpus – news broadcasts in *MPEG-1 Audio Layer II* versus read speech in 16kHz *WAVE* format. We need to convert them into a single format for training of the recognizer.

Audio tracks from the *MPEG-1* video files of the news stories are extracted and converted to 16kHz mono-aural *WAVE* format. They are then used to retrain the seed acoustic models in the base syllable recognizer, which is Hidden Markov Model-based (HMM-based) and uses acoustic models for syllable initials (3-state HMMs) and syllable finals (5-state HMMs). These models are right content-dependent HMMs with 16 Gaussians mixtures. The acoustic features used are 12 mel-frequency cepstral coefficients (MFCC) with the log energy and augmented with the first and second derivatives (39 parameters per input vector).

The seed acoustic models are further retrained using down-sampled audio news data. 2.47 hours of hand-transcribed news data are blindly segmented into 20-second segments for embedded training. The phonetic representations used follows the standard defined by the Linguistic Society of Hong Kong (LSHK) [57]. The pronunciations in the models are extracted from a 41k-word lexicon (CULEX^{TM}) and a 10k-word lexicon (CUPDICT^{TM}) [58].

Evaluation is based on another 2.75 hours of hand-transcribed news data. The base syllable accuracy of our recognizer is found to be 44.4%. The low accuracy is mainly due to harsh acoustic conditions (especially for audio record-

ings from the field speech) 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, reporter and interview (i.e. field) speech. Syllable accuracies are shown in Table 3.1. 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.

	Studio speech	Field speech	
	Anchor	Reporter	Interviewee
Syllable accuracies	59.3%	43.3%	27.0%
		39.2% (Overall)	

Table 3.1: Base syllable accuracies of audio indexing by base syllable recognition. Anchor speech is clearly articulated and recorded in the studio with favorable ambient conditions. Reporter and interviewee speech are spontaneous and recorded from the field, possibly with harsh acoustic conditions.

Retained acoustic models are used for indexing of the audio tracks of the news collection. The recognized syllables generated are used for indexing of the Cantonese spoken documents.

3.2.2 Indexing Units

As explained in Section 1.2, the use of bigrams and skipped bigrams in retrieval can contribute to problem of Chinese word tokenization and resolve ambiguity in Chinese homophones. Previous experiments in Cantonese spoken document retrieval [23] [59] have shown that overlapping character/syllable bigrams are effective units for indexing and retrieval. Therefore, we use overlapping character/syllable bigrams and skipped overlapping character/syllable bigrams to

capture Chinese words and verbalized Chinese abbreviations. Table 3.2 shows procedure for forming different indexing units.

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/

Table 3.2: Procedure for forming text-converted overlapping syllable bigrams and skipped bigrams.

In Table 3.2, the Chinese word in the first row refers to the Chinese University of Hong Kong. We can segment the four Chinese characters into two words, 中_文 (meaning: Chinese) and 大_學 (meaning: university) by means of word tokenization. These two words can be captured in the second row using overlapping bigrams. In the formation of overlapping skipped syllable bigrams (see row four), it captures the abbreviation 中_大 of the Chinese University of Hong Kong.

3.3 Known-Item Retrieval Task

Recall that each audio document in the corpus has a corresponding textual summary with a title (see Figure 2.2). However, the news stories in the Cantonese news corpus are not classified into topics and the corpus does not contain relevance judgments. Hence we formulated a known-item retrieval (KIR) task for the Cantonese SDR experiments. We used the summary title as a query and the rest of the summary as textual document. The query is used to retrieve its corresponding textual or spoken document from the pool. The goal of KIR is to generate single relevant document for each query. Figure 3.2

illustrates the idea of KIR in this work. According to the similarities between the query and documents, KIR task can rank the retrieved documents and a rank will be given to the relevant document.

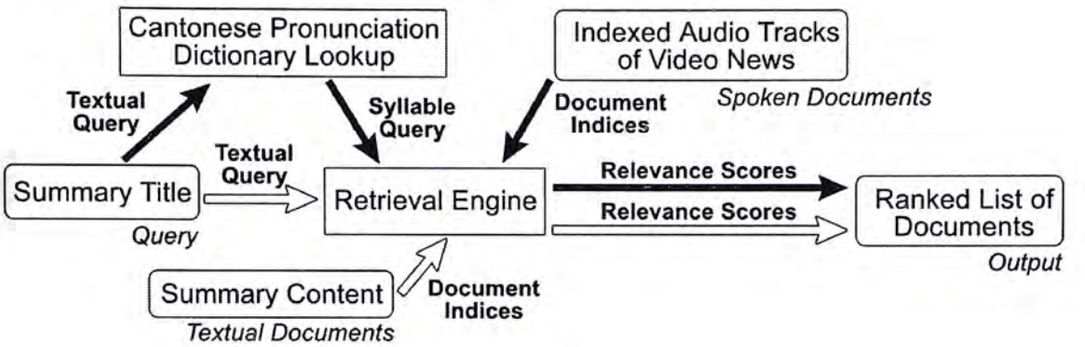

Figure 3.2: An illustration of the KIR task for Cantonese SDR.

3.3.1 Evaluation – Average Inverse Rank

Evaluation is based on the ranked list of retrieved documents from the retrieval engine. Recall that each query only has one relevant document in KIR. We check the ranks of the relevant documents for the given queries. For performance measure, we average the inverse of the ranks of retrieved documents. This measure is known as average inverse rank (AIR) / mean reciprocal rank and is expressed as shown in Equation 3.1. High values in AIR indicate good retrieval performance. Perfect retrieval will produce $AIR = 1$ since all queries have their relevant documents ranked first in all the retrieved lists. If the relevant documents are consistently ranked very low in the retrieved lists, the retrieval performance is low and the valule of AIR is closed to zero.

$$AIR = \frac{1}{N_d} \sum_{i=1}^{N_d} \frac{1}{rank_i} \tag{3.1}$$

where N_d is the total number of query-document pairs (the total number of documents in the collection) and $rank_i$ is the rank of the i^{th} document retrieved using query i.

3.4 Retrieval Model

The objective of SDR is to retrieve spoken documents to meet user's requests. In this work, we use a statistical model – vector-space model (VSM) [60], to perform all retrieval experiments.

Each document in the collection is represented in the form of a vector d. We weigh each indexed term in all document vectors by a factor, which indicates the importance of the corresponding term. We use the raw frequency of a term in a vector to be its weight and refer it as term frequency (TF). The weight of a term is proportional to its occurrence in a specific document vector. The document vector d can be written as:

$$d[i] = \ln(tf_d[i]) + 1.0 \tag{3.2}$$

where $tf_d[i]$ is term frequency of term i in document vector d.

The user's query is also represented as a query vector q. Again, we weigh each indexed term in the query vector by a TF factor, which reflects the importance of the corresponding query term. In addition to the TF, we use the inverse document frequency (IDF) to reflect the degree of discrimination of a query term in document vectors. Terms that appear in many documents are not very useful in discriminating between of relevant and non-relevant documents. Common words have high value in the TF. The IDF can balance the adverse effect of common words in document vectors. IDF is given by:

$$IDF[i] = \ln(\frac{N_d + 1}{n_i}) \tag{3.3}$$

where n_i is the number of documents with term i.

In our experiments, we combine TF with IDF to be the weighting function for query vector:

$$q[i] = [\ln(tf_q[i] + 1.0)] \times \ln(\frac{N_d + 1}{n_i}) \tag{3.4}$$

where $tf_q[i]$ is the TF of term i in query vector q.

The similarity between a document and a query vector is calculated as the inner product of the two vectors. Long documents contain more terms and more times. VSM tends to favor long documents and tends to increase the similiarity score. We use cosine normalization (CN) to reduce the adverse effect of term repetition in long documents. Cosine normalized similarity / retrieval score is defined as:

$$Similarity_{cosine}(q, d) = \frac{q \cdot d}{\|q\|\|d\|} \tag{3.5}$$

We can rank all the documents in the collection according to the retrieval scores calculated. The retrieval engine can produce a ranked retrieval list of N documents for each query.

3.5 Experimental Results

Table 3.3 shows the speech retrieval performance based on overlapping character bigrams/skipped bigrams, text-converted syllable bigrams/skipped bigrams and recognized syllable bigrams/skipped bigrams. Overlapping character bigrams are derived from textual summary prose. Overlapping text-converted syllable bigrams are syllables bigrams converted from the character bigrams by pronunciation lookup. This setup provides a benchmark performance comparable to the case of perfect syllable recognition for indexing the Cantonese spoken documents. Overlapping recognized syllable bigrams are derived from syllable recognition of the audio documents. The is the setup for actual Cantonese SDR.

Retrieval unit	**AIR**
Overlapping character bigrams	0.977
Overlapping character bigrams and skipped bigrams	0.976
Text-converted syllable bigrams	0.973
Text-converted syllable bigrams and skipped bigrams	0.974
Recognized syllable bigrams	0.612
Recognized syllable bigrams and skipped bigrams	**0.633**

Table 3.3: Retrieval performances based on average inverse rank using overlapping character bigrams/skipped bigrams and text-converted syllable bigrams/skipped bigrams.

The high retrieval results using character bigrams/skipped bigrams suggested that the summary titles can succinctly capture the key terms in the news story. Retrieval performance using overlapping character bigrams and skipped bigrams is slightly lower that of character bigrams is due to the generation of *extra* words by overlapping skipped bigrams. For example, the fourth row of Table 3.2 shows that skipped bigrams can capture the common abbreviation 中_大 of the Chinese University of Hong Kong. However, at the same time, it also generated the word 文_學 (meaning: Literature), which is not directly related to the university. We also observe an overall degradation of retrieval performance due to the imperfect recognition after comparison between results using text-converted syllable (row four) and recognized syllable (row six).

3.6 Chapter Summary

In this chapter, we present the background information on our work in Cantonese spoken document retrieval. Each news story in the collection is ac-

companied by a textual summary with a title. We formulated a known-item retrieval task for retrieval experiments. The summary title is used as a query to retrieve its corresponding textual and spoken document. We indexed the spoken documents (audio tracks of the video news) using a Cantonese base syllable recognizer. Retrieval experiments are performed using overlapping character/syllable bigrams and skipped bigrams. Retrieval experiments are performed using character bigrams and skipped bigrams. Word-based text retrieval obtained a retrieval performance with *AIR=0.976*. The use of text-converted syllable bigrams and skipped bigrams has a comparable performance with *AIR=0.974*, which is an approximation on perfect recognition. Results based on recognized syllable bigrams and skipped bigrams gave *AIR=0.633*. A possible reason of the severe degradation of performance is the speech recognition errors during indexing.

Chapter 4

The Use of Audio and Video Information for Monolingual Spoken Document Retrieval

This chapter reports our initial study in automatic story segmentation and automatic extraction of anchor speech. Automatic story segmentation [61] [62] by means of video-based segmentation algorithm is used to replace hand-segmentation of news stories from news programs in our video archive. Results in Table 3.1 indicate that speech recognition performance for audio indexing is more reliable for the anchor speech from the studio, in comparison with the reporter/interviewee speech from the field. In order to reduce the adverse effect of speech recognition errors and the audio indexing effort required, we have devised three methods for the automatic extraction of anchor speech [63]. They are (i) video-based segmentation, which utilizes video frame information only, (ii) audio-based segmentation, which utilizes audio information only and (iii) fusion of video- and audio-based segmentation, which fuses both audio and video information for extraction. All of the methods aim to locate the studio-to-field transition in the news data.

The temporal syntax of the news video from the local television station is rather straightforward – it follows the template of as shown in Figure 4.1. A news story typically begins with anchor shot[4]and followed by one or more field shots. Studio-to-field transition is the shot boundary between an anchor shot and a field shot. With the information of studio-to-field transition, we can segment the audio track for each news story into the segment of anchor speech and the segment of live footage. Thereafter, these segments can be processed individually for speech retrieval [64]. Observation shows that average duration of anchor speech is only a quarter of news story, processing of anchor speech only can reduce the indexing effort. The information of field-to-studio transition (i.e. shot boundary between a field shot and an anchor shot) from the algorithm can be used for the purpose of automatic story segmentation.

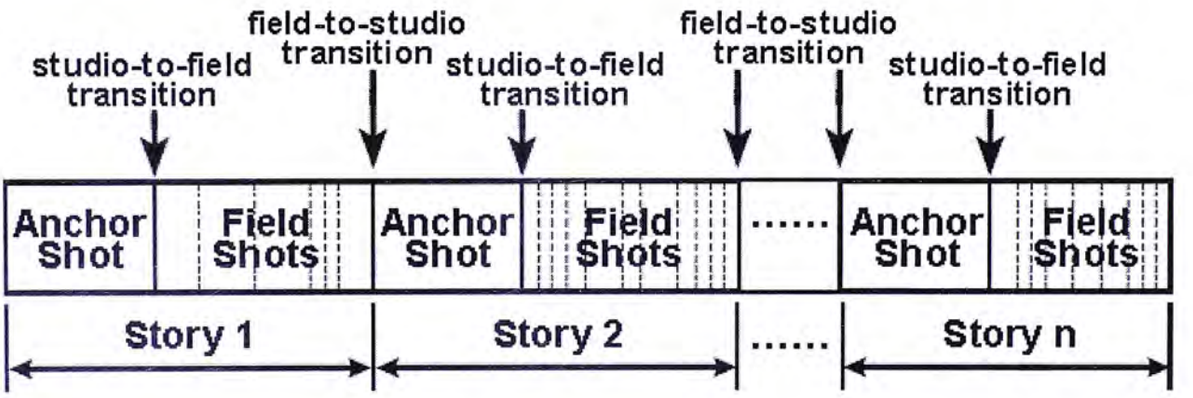

Figure 4.1: A simplified template of the news programs collected.

4.1 Video-based Segmentation

We adopted the video-based segmentation algorithm developed by [65]. It consists of four modules as shown in Figure 4.2. Each module will be explained

[4]A shot is the basic structuring element of video. A shot is a contiguous recording of one or more video frames depicting a contiguous action in time and space. During a shot, the camera may remain fixed, or may exhibit such motions as panning, tilting, zooming, tracking, etc.

in detail in the following sections.

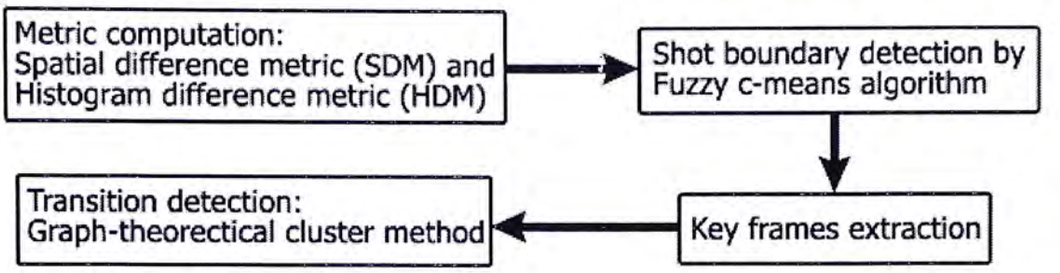

Figure 4.2: Control flow of the video-based segmentation algorithm.

4.1.1 Metric Computation

We have extracted the video frames from the *MPEG-1* video files of the news programs. *MPEG-1* is used because the video files in the *MPEG-1* format have higher video frames quality than in the *RealMedia*™format.[5] We have sampled one out of every five video frames from the *MPEG-1* files to obtain a sparser frame sequence. A sparser frame sequence is used instead of a continuous one, so as to reduce the computation cost. Then we compute the gray level (intensity histograms) and the color histograms of every pair of consecutive frames in this sequence. Figure 4.3 illustrates that if both frames are from the same anchor shot, they tend to have very similar histograms. However, if the pair of frames crosses a shot boundary, their histograms are very different, as illustrated in Figure 4.4. This includes the cases of (i) one of the consecutive frames belongs to an anchor shot while another a field shot, (ii) both of the consecutive frames are from different field shots or (iii) both of the consecutive frames are from different anchor shots. In order to compare the qualitative difference between a pair of consecutive frames, we use two metrics – the his-

[5]Video files in the *MPEG-1* format have frame size and frame rate of 352 × 288 pixels and 25 fps. Video files in the *RealMedia*™format have frame size and frame rate of 160 × 112 pixels and 15 fps.

togram difference metric (HDM) and the spatial difference metric (SDM), as shown in Equations 4.1 and 4.2. HDM is used to compare gray level (intensity histograms) between two frames. The comparison in HDM is obtained by counting the number of pixels with color k in each frame. However, HDM may be insensitive to small changes / camera movements. Hence SDM is used to capture frames with little change. SDM is a pixel-wise comparison of the intensity between two frames.

$$HDM = \frac{1}{M \times N} \sum_{k=1}^{L} |H_t(k) - H_{t+5}(k)| \tag{4.1}$$

$$SDM = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} |I_t(i,j) - I_{t+5}(i,j)| \tag{4.2}$$

where $M \times N$ is the frame size, which is 352×288 in this work,

$I_t(i,j)$ denotes the intensity of a pixel at location (i,j) in the t^{th} frame,

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

L is the total number of colors.

If we plot the SDM or HDM values against the frame pair number, we obtain a sequence of pulses as shown in Figures 4.5 and 4.6. High values often result from large content changes in the frames and are indicative of consecutive frames that crosses a shot boundary.

4.1.2 Shot Boundary Detection

We normalize the SDM and HDM values using the following equations,

$$SDM_{normalized} = \frac{SDM - SDM_{minimum}}{SDM_{maximum} - SDM_{minimum}} \tag{4.3}$$

$$HDM_{normalized} = \frac{HDM - HDM_{minimum}}{HDM_{maximum} - HDM_{minimum}} \tag{4.4}$$

For each consecutive frame pair in our frame sequence, we plot its normalized SDM and HDM values in a scatter plot as shown in Figure 4.7.

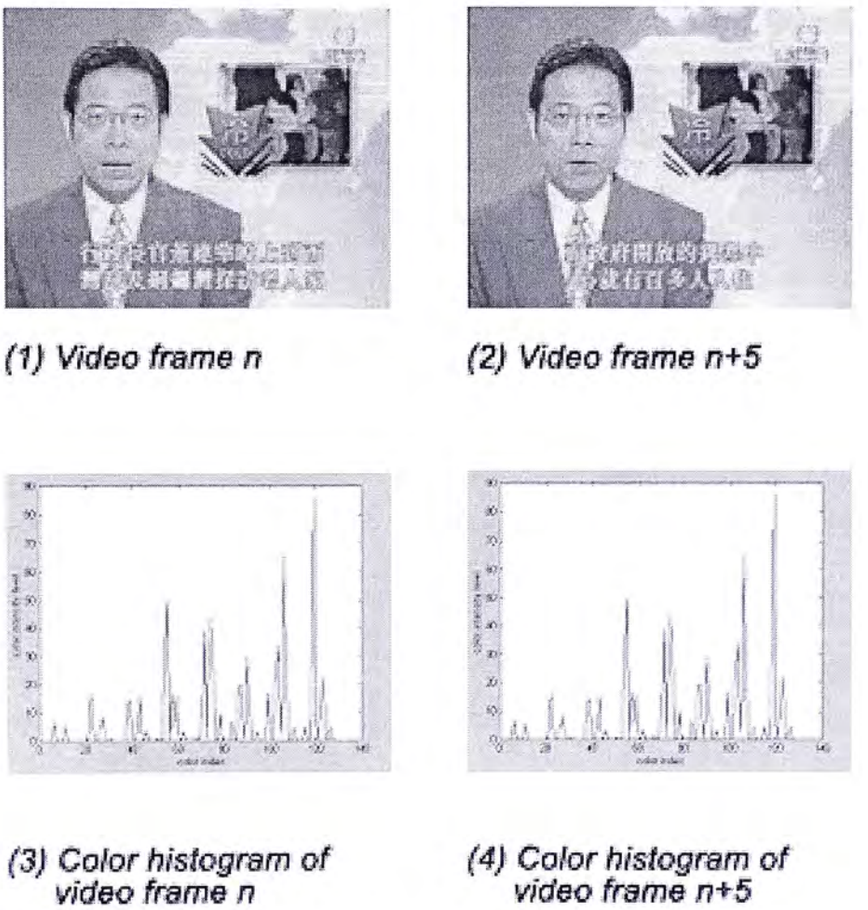

Figure 4.3: Color histograms of the video frames in the same anchor shot.

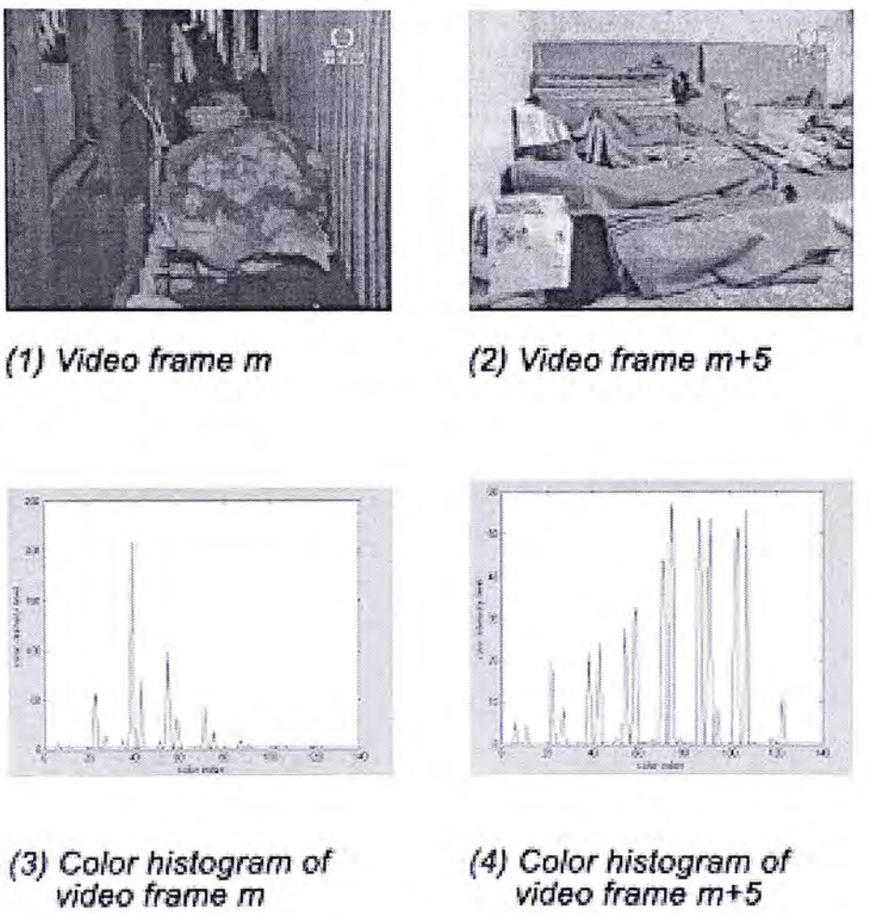

Figure 4.4: Color histograms of the video frames across a shot boundary for field shots.

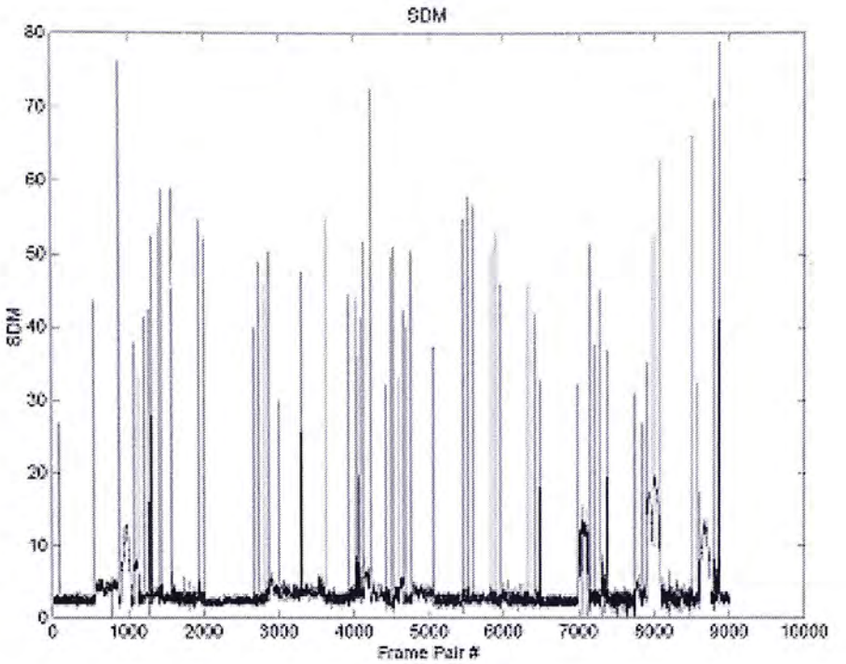

Figure 4.5: A plot of the SDM against frame pair number.

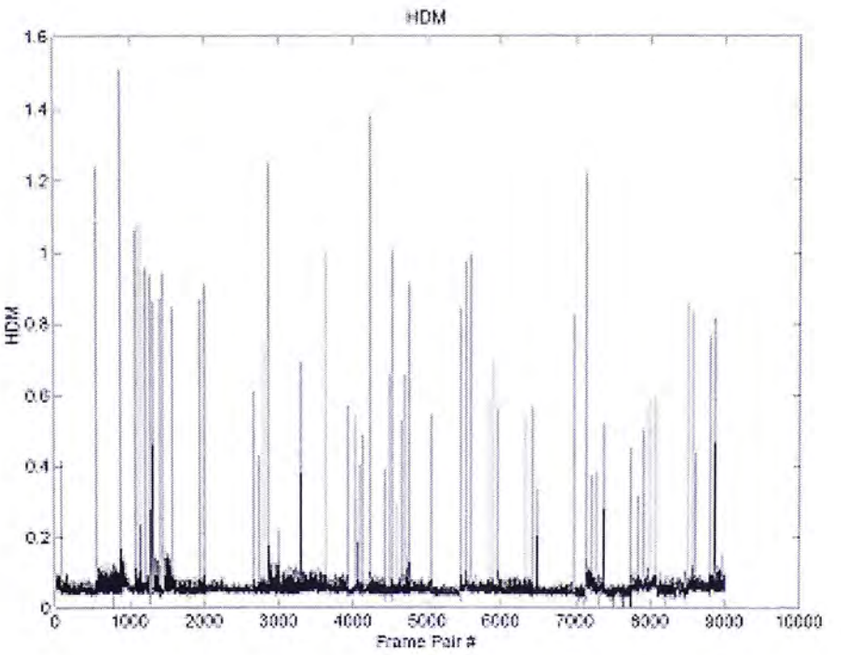

Figure 4.6: A plot of the HDM against frame pair number.

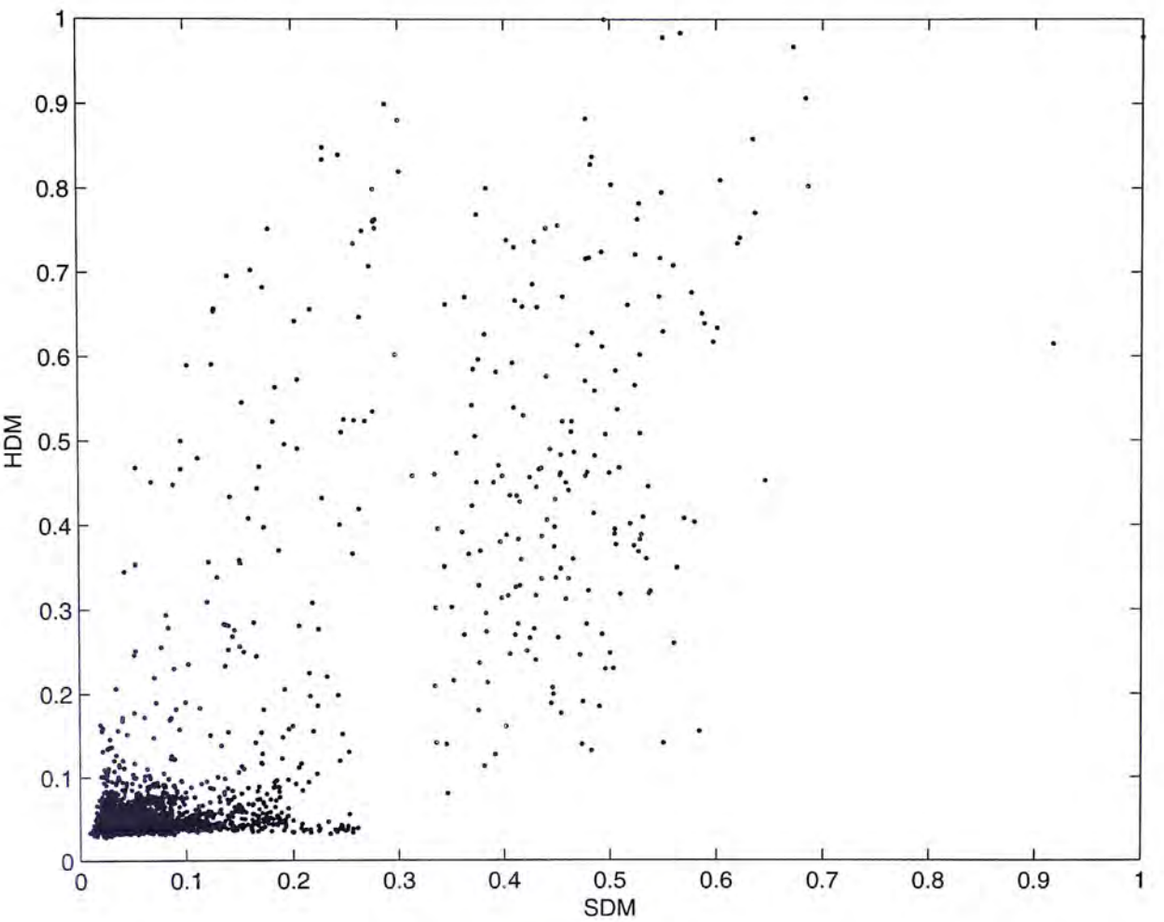

Figure 4.7: A plot of the normalized HDM against SDM.

To detect the frame pairs with significant change, we partition this feature space into two subspaces, the significant change (SC) category and non-significant change (NSC) category. Hard clustering assumes there are well-defined boundaries between the subspaces. Hence hard clustering assigns each data point (feature vector) to one and only one of the subspaces, with a degree of membership equal to either zero or one. This model does not reflect the description of real data, where boundaries between subspaces might be fuzzy – i.e. the new shot may fade in, the previous shot can fade out, or we have a combination of both fade in and fade out. Hence we use the unsupervised fuzzy c-means (FCM) clustering algorithm [66] to classify the feature vectors into SC and NSC categories based on an objective function. The objective function represents the inner product metric (distance measure) from any given feature vector $F_D(t)$ to a centroid v. The distance is weighted by the membership degree of that feature vector u_{it} and the fuzziness of the cluster m. Using FCM, the degree of membership of a feature vector to a category can be any value between zero to one. The objective function $J_m(U, v)$ and the fuzzy 2-partition space M_{f2} for feature vectors are defined as shown in Equations 4.5 and 4.6. The FCM algorithm is iteratively updating the centroids and the membership degrees for each feature vector. The iteration is based on minimizing the defined objective function. By minimizing the objective function, we can obtain the optimal fuzzy space partition U^* and the optimal cluster prototype v^*. In our application, the FCM algorithm starts with an initial guess for the centroids that are intended to mark the mean location of each cluster. We have set the value of fuzziness m to be 2 in our experiments. The value 2 is chosen based on performance comparison experiments of the FCM algorithm with $m = 2$, 3 and 4. The results showed that the FCM algorithm performed best when $m = 2$. Two stopping criteria have been applied on the FCM algorithm during implementation. They are the maximum number of iterations and the

minimum amount of improvement. We have arbitrarily set their values to 100 and 10^{-6}. The algorithm will stop if the number of iterations is geater than 100 and/or the amount of improvement in $J_m(U, v)$ is smaller than 10^{-6}.

$$J_m(U, v) = \sum_{t=1}^{T} \sum_{i=1}^{2} (u_{it})^m \cdot \|F_D(t) - v(i)\|^2 \tag{4.5}$$

$$M_{f2} = \left\{ \begin{array}{c} U \in V_{2T} | u_{it} \in [0, 1] \ \forall i, t; \\ \sum_{i=1}^{2} u_{it} = 1 \ \forall k; \ 0 < \sum_{t=1}^{T} u_{it} < T \ \forall i \end{array} \right\} \tag{4.6}$$

where $U \in M_{f2}$;

$v(i) \in R^{2\times 2}$ is the centroid of fuzzy subset u_i, $i = 1, 2$;

$m \in [1, \infty)$ is the weighting exponent to control the "fuzziness"of the resulting clusters ($m = 2$ in this work);

u_{it} denotes the degree of membership of the t^{th} vector $F_D(t)$ belonging to the i^{th} category;

V_{2T} is the set of real $2 \times T$ matrices; and

T is the total number of feature vectors.

The two clusters output from FCM algorithm is shown as Figure 4.8. The centroids of the clusters in the last iteration are indicated with the large 'O' and 'X'. There are a few feature vectors with "fuzzy"results at the boundary between the two clusters. These feature vectors are labeled with $\otimes$ in Figure 4.8.

We then implemented the defuzzifying operation on the classification result by the FCM algorithm to get the *crisp* classification result as shown in Figure 4.9. The defuzzifying operation rules are defined in Equations 4.7 and 4.8. They are used to compare the degree of membership of a feature vector to both of the categories. If the degree of membership of a vector to the SC category u_{1t} is equal to or greater than that of the NSC category u_{2t}, the vector will be classified into the SC category and vice versa.

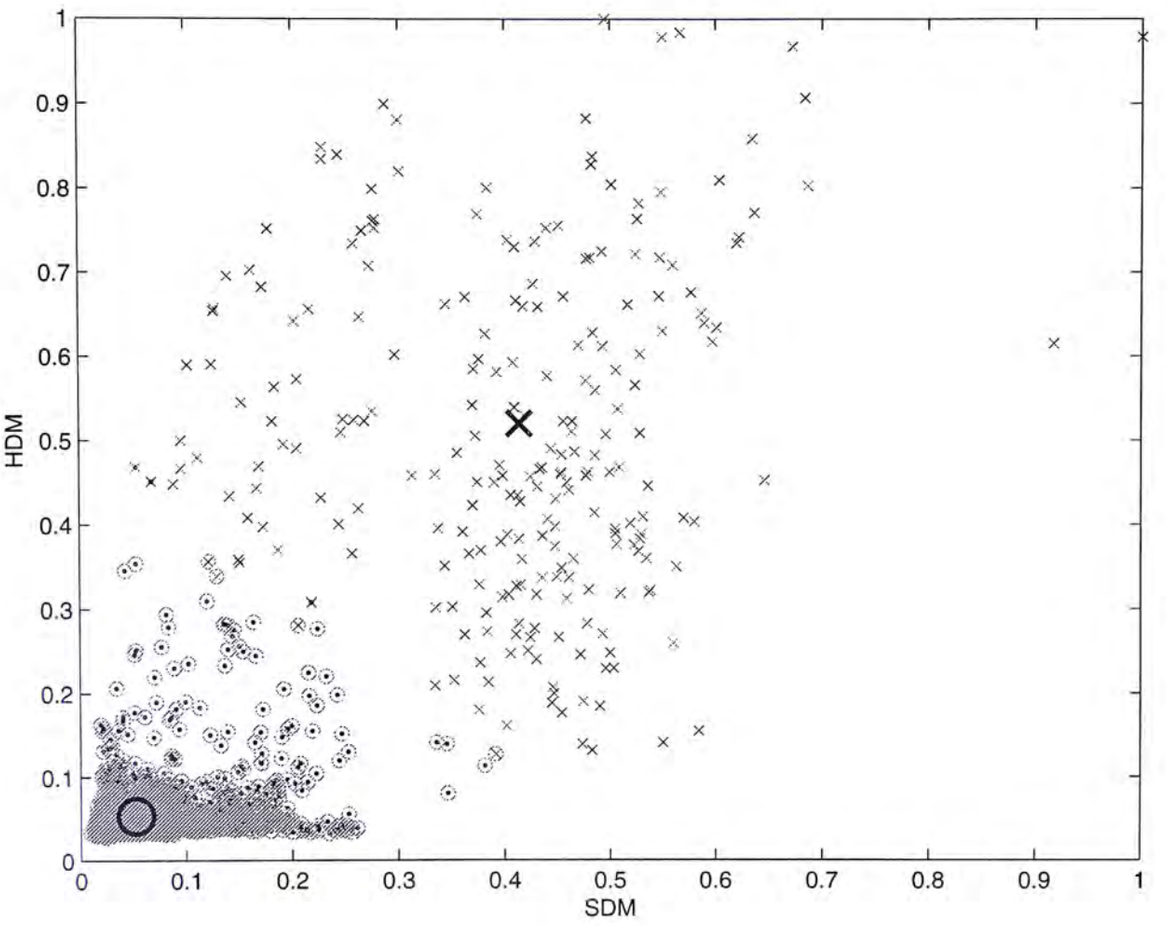

Figure 4.8: Two clusters are formed after FCM based on the input as shown in Figure 4.7.

$$F_D(t) \in SC, \text{ if } u_{1t} \geq u_{2t}; \text{ and} \tag{4.7}$$

$$F_D(t) \in NSC, \text{ if } u_{2t} > u_{1t} \tag{4.8}$$

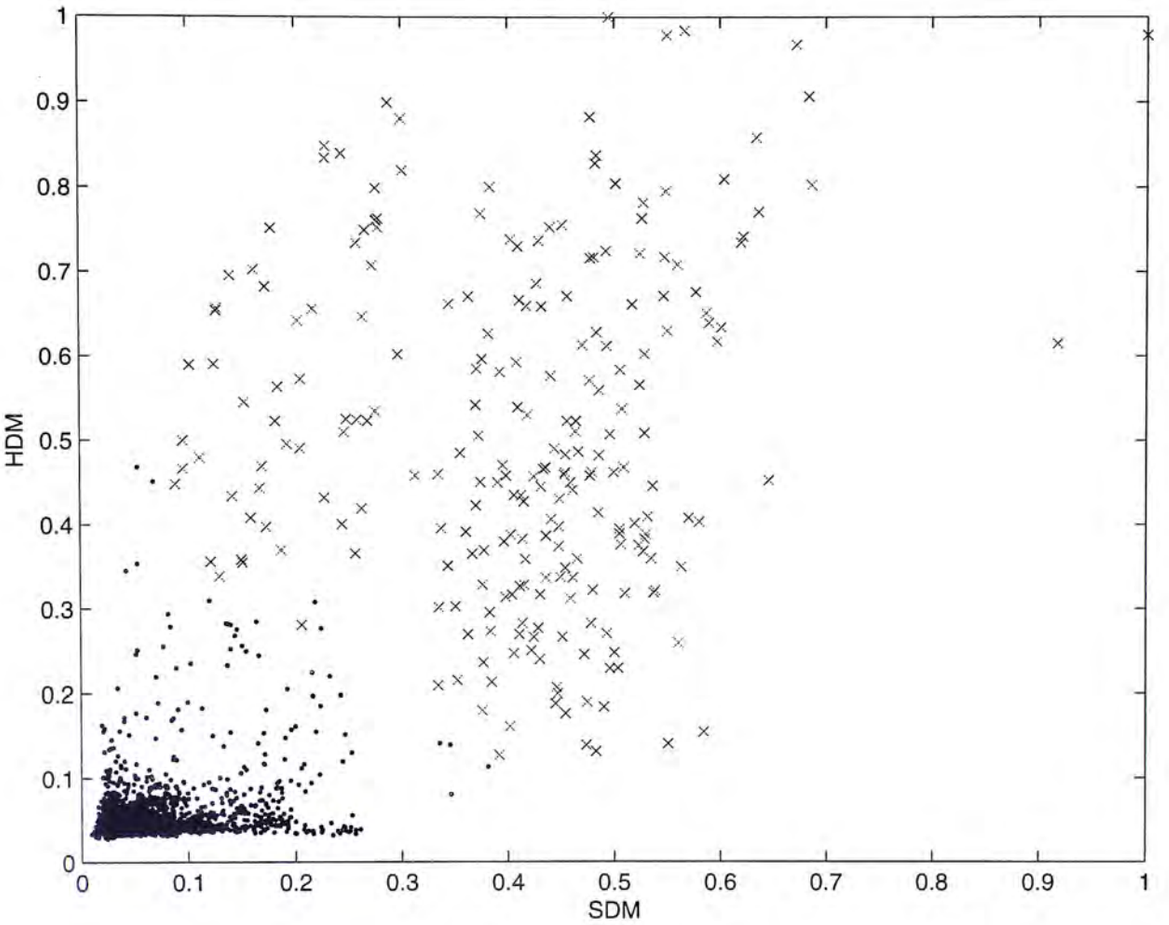

Figure 4.9: The classification result of Figure 4.7. Frame pairs with significant change labeled as shot boundaries are indicated with 'x'.

According to the temporal structure of television news program, frames between two consecutive boundaries form a shot. We can segment a news program into individual shots using the shot boundaries, i.e. the feature vectors (with their corresponding frame numbers) in the SC category.

The next step is to classify the shots into anchor shots or field shots. For simplication, we extract the first frame from each shot for further classification.

The frame extracted will be the representative of that shot and is the key frame of that shot.

4.1.3 Shot Transition Detection

We observe that key frames from anchor shots follow the four patterns of anchor frames in the video archive (see Figure 2.3). There are at least two key frames in each of the pattern in the news program. Most key frames of field shots are very different from one another. Since the background region of the anchor key frames tends to be relatively fixed, the different studio key frames with the same pattern generally have similar color histograms. 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) analysis algorithm [67].

Key frames in the 128-dimensional color histogram space are illustrated as the nodes in Figure 4.10. We formed a minimum spanning tree (MST) using a single link algorithm to link all the key frames in this space. The path length connecting any two nodes (representing two key frames) in MST is proportional to the difference between the HDM of the two key frames. We then sever the edges of the tree, which have a distance larger than a threshold. There will be four connected clusters remained as shown in Figure 4.11. Each of these clusters contains key frames that closely resemble one another. We have four clusters corresponding to the four types of anchor shots in Figure 2.3. Hence the key frames in these clusters are treated as studio shots (i.e. field-to-studio transitions / story boundary). The other frames (singletons) are automatically treated as field shots (i.e. studio-to-field transitions).

According to the temporal syntax of a news program, we can use the story boundary's frame number (field-to-studio transition) to segment the program into individual news stories. A news story is typically begun with anchor shot

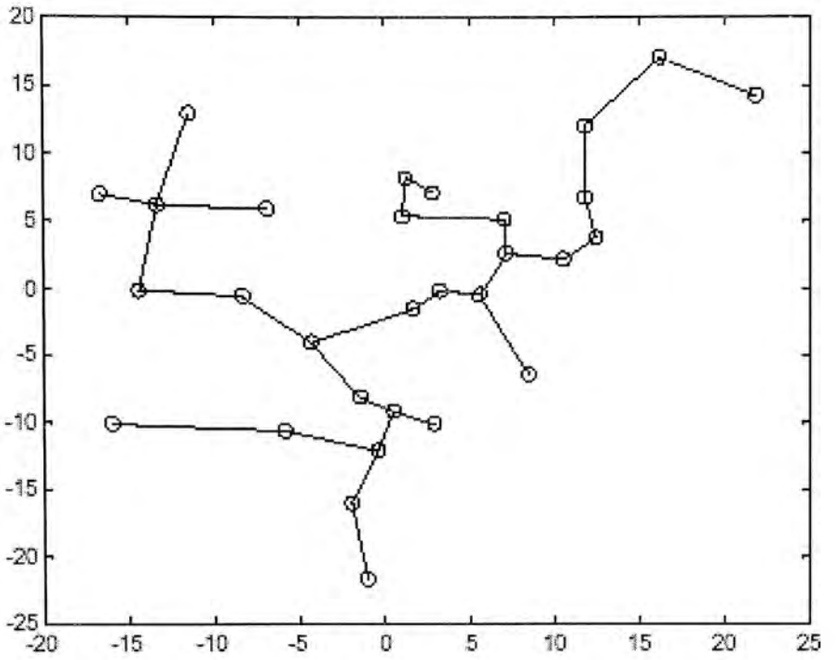

Figure 4.10: A plot of minimum spanning tree.

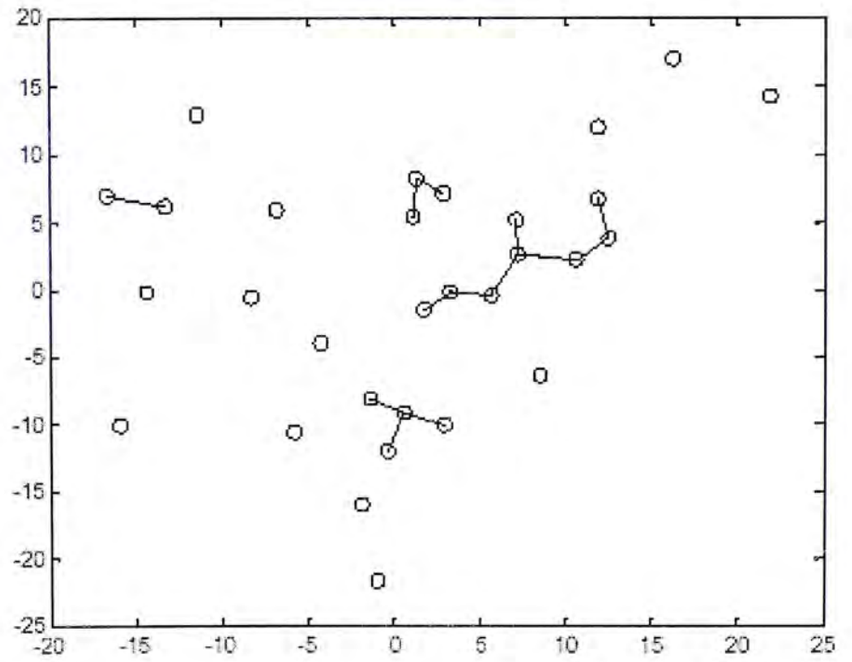

Figure 4.11: The remaining clusters of Figure 4.10 after deleting the edges with distance larger than a threshold.

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

4.2 Audio-based Segmentation

In this work, we made another attempt to extract anchor speech based only on the audio information. This method aims to capture differences in the acoustic signal since anchor speech tends to have little noise, while field speech may contain music, environmental noises, etc.

4.2.1 Gaussian Mixture Models

We use single-state Gaussian Mixture Models (GMM) [68] [69] for audio-based segmentation. We trained one GMM to be the studio model and another to be the field model (see Figures 4.12 and 4.13) by applying the Baum-Welch algorithm (also known as the forward-backward algorithm) on five hours of audio data (a subset of the 60.4 hours mentioned in Table 2.3 from our corpus). The number of Gaussian mixtures was increased exponentially from 1 to 64 during the training stage. At 64 mixtures, the GMM can correctly extract around 90% of the anchor/studio speech segments from the five hours of training data.

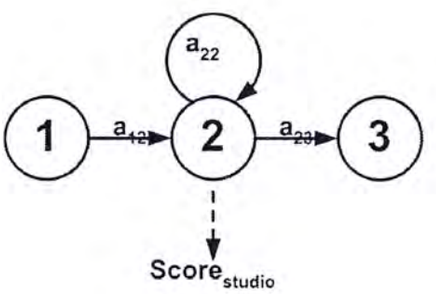

Figure 4.12: A simple GMM anchor model with 3 states.

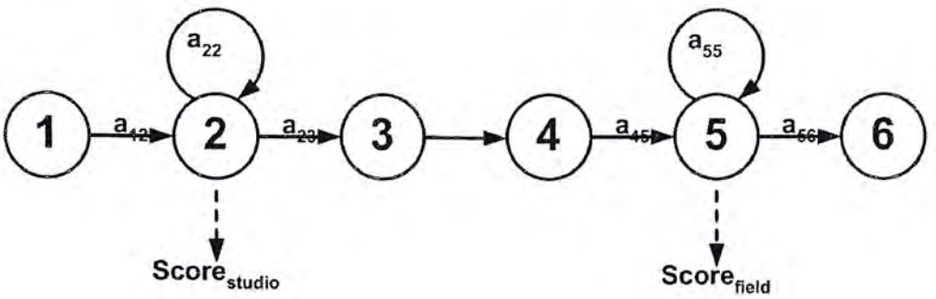

Figure 4.13: A simple GMM model for studio-to-field transition detection. An anchor model and a field model merge together to form the GMM model with 6 states.

4.2.2 Transition Detection

After determination of the number of mixtures used in the GMM, we used these GMM to process the entire audio data set (60.4 hours). The use of GMM aims to distinguish news stories without field shots from those with studio-to-field transitions. We first compute the cumulative score for a news story with T speech frames by traversing with the studio model only:

$$\begin{aligned} Score_{studio\_only} &= Score_{studio} \\ &= \prod_{t=1}^{T}[\sum_{i=1}^{64} w_i \cdot N_{studio}(j_t; \mu_i, \sigma_i)] \end{aligned} \tag{4.9}$$

under the conditions:

$$w_i = \frac{\sum_{j=1}^{M} I_i(j)}{\sum_{j=1}^{M}\sum_{k=1}^{64} I_k(j)} \tag{4.10}$$

$$\sum_{i=1}^{64} w_i = 1 \tag{4.11}$$

where w_i are the mixture weights for the Gaussians that are obtained from the training set;

M is the number of observations in the training set;

$I_i(j)$ is an indicator function, $I_i(j) = 1$ if an observation j is associated with the mixture component i in the training set; and

$N_{studio}(\cdot; \mu_i, \sigma_i)$ is obtained from the studio model for the mixture component i.

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

$$\begin{aligned} Score_{studio\_to\_field} &= Score_{studio} \cdot Score_{field} \qquad (4.12) \\ &= (\prod_{t=1}^{T_t} [\sum_{i=1}^{64} w_i \cdot N_{studio}(j_t; \mu_i, \sigma_i)])(\prod_{s=T_t}^{T} [\sum_{k=1}^{64} w_j \cdot N_{field}(j_s; \mu_k, \sigma_k)]) \end{aligned}$$

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. Figure 4.14 shows a sample output of the stories with filename 1999070711 and 1999070712, from the audio segmentation algorithm. The results are in the format of "segment_start_time segment_end_time shot_notation log_likelihood". "segment_start_time" and "segment_end_time" are in the unit of $10^{-7} seconds$.

Filename: 1999070711

0 148719997 anchor -62.017597

.

Filename: 1999070712

0 *209039993* anchor -62.496258

209040000 796079986 other -65.450096

.

Figure 4.14: A sample output from the audio segmentation algorithm. The algorithm detected that the news story with filename 1999070711 is consisted of anchor speech only. The news story with filename 1999070712 has a studio-to-field transition at 20.9 seconds (i.e. 209039993×10^{-7} seconds).

4.3 Performance Evaluation

For evaluation purpose, we have manually labeled all the studio-to-field and field-to-studio transitions (story boundaries) in our 60-hour video corpus. The transitions are labeled based on video information. The manual annotations indicate that 1,545 of the news stories (around 95%) contain a single studio-to-field transition and the remaining news stories do not have field shots. Hence the evaluation is based on 1,545 stories only. We consider a 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 two seconds of audio).

In general, the performance of segmentation algorithms are measured based on precision and recall. Precision refers to how many segmentation outputs are correct and is defined as the ratio of the number of transitions detected correctly to the total number of transitions detected. Precision value falls into the range from zero to one.

$$Precision = \frac{number\ of\ transitions\ detected\ correctly}{number\ of\ transitions\ detected\ by\ the\ algorithm} \tag{4.13}$$

Recall measure the performance of an algorithm by finding the ratio of correctly detected transitions to the total number of transitions in the experimental archive.

$$Recall = \frac{number\ of\ transitions\ detected\ correctly}{number\ of\ transitions\ in\ the\ collection} \tag{4.14}$$

4.3.1 Automatic Story Segmentation

From the 60-hour video corpus, the automatic story segmentation algorithm labeled 1,431 story boundaries. Of these, 1,335 are correct. Hence automatic story segmentation achieved a precision of 0.933 and a recall of 0.864.

Figure 4.15 shows the results of automatic story segmentation by means of video-based segmentation.

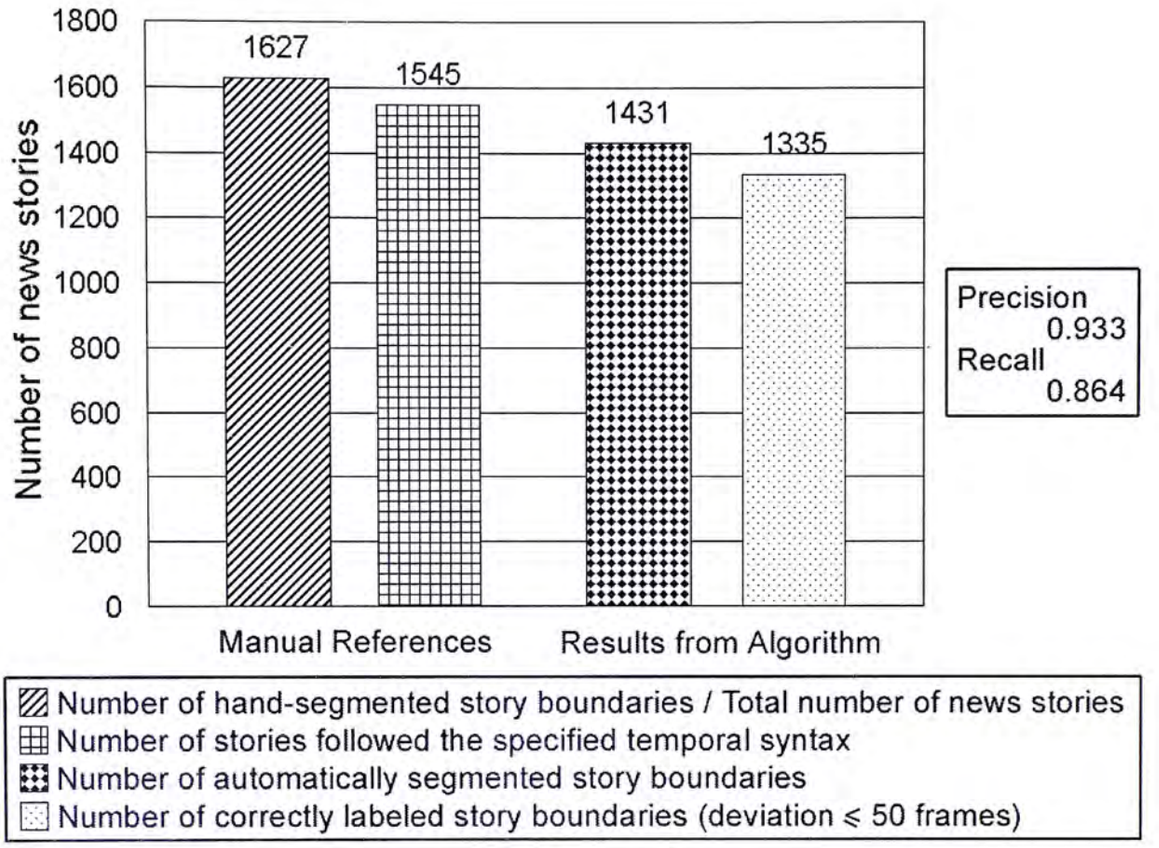

Figure 4.15: Results of the automatic story segmentation algorithm by means of video-based segmentation.

4.3.2 Video-based Segmentation Algorithm

We applied the video-based segmentation algorithm on all the 1,627 video files in our corpus. Our video-based segmentation algorithm labeled 1,431 news stories with a studio-to-field transition and the remaining stories were labeled with zero transition. Evaluation shows that 1,365 of the automatically detected transitions correspond to the manual ones. Hence video-based segmentation algorithm, for the extraction of anchor/studio speech segments, achieved a precision of 0.954 and a recall of 0.884 (see Table 4.1).

4.3.3 Audio-based Segmentation Algorithm

We evaluate the 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 4.1 together with those from the video-based segmentation algorithm. It should be noted that manual annotation is based on video frames and we have found 306 news stories (about 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 (see category (ii) in Figure 3.1). 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. Table 4.1 summarized the performance of the video- and audio-based segmentation algorithms. 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.

	Video-based	**Audio-based**
Number of transitions labeled by algorithm	1,431	1,376
Number of transitions labeled correctly (deviation $\leq$ 50 frames)	1,365	1,208
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 4.1: 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.

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 where 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 a recall of 0.627 within this special subset of news stories (see Table 4.2).

Number of stories contain transition in video track only	306
Number of stories labeled with zero transition in audio-based segmentation algorithm	251
Number of stories labeled correctly	192
Precision	0.765
Recall	0.627

Table 4.2: Results of the audio-based segmentation algorithm on the special subset of news stories after further investigation.

4.4 Fusion of Video- and Audio-based Segmentation

We have found that there is a special subset of news stories that only have anchor-to-field transitions in video. Based on the special feature of these stories, we devised the third method for automatic extraction of anchor/studio speech. This method fuses results from video-based segmentation with those from audio-based segmentation to further improve automatic location of studio-to-field transitions. Table 4.3 shows statistics relating to the presence/absence of studio-to-field transitions in the audio and video tracks of our news stories: Based on these statistics we have devised the following fusion strategy:

Case 1: Both video- and audio-based segmentation algorithms detect studio-

	Transition in audio	No transition in audio
Transition in video	**1,239** (category (i))	**306** (category (ii))
No transition in video	**0**	**82** (category (iii))

Table 4.3: 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 1,627. Illustration of the categories are shown in Figure 3.1.

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 4.1).

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 4.3).

Case 3: Both video- and audio-based segmentation do not detect any transition – the entire audio track is used in spoken document retrieval.

Case 4: Only the audio-based segmentation detects a studio-to-field transition – the entire audio track is used in spoken document retrieval since no such category of news story should exist (see Table 4.3).[6]

4.5 Retrieval Performance

Figure 4.16 shows the retrieval results for various methods of extracting the anchor/studio speech segments. The results without using any extraction method (entire audio track is used) and manual video-based segmentation are included as references (rows 1 and 2).

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

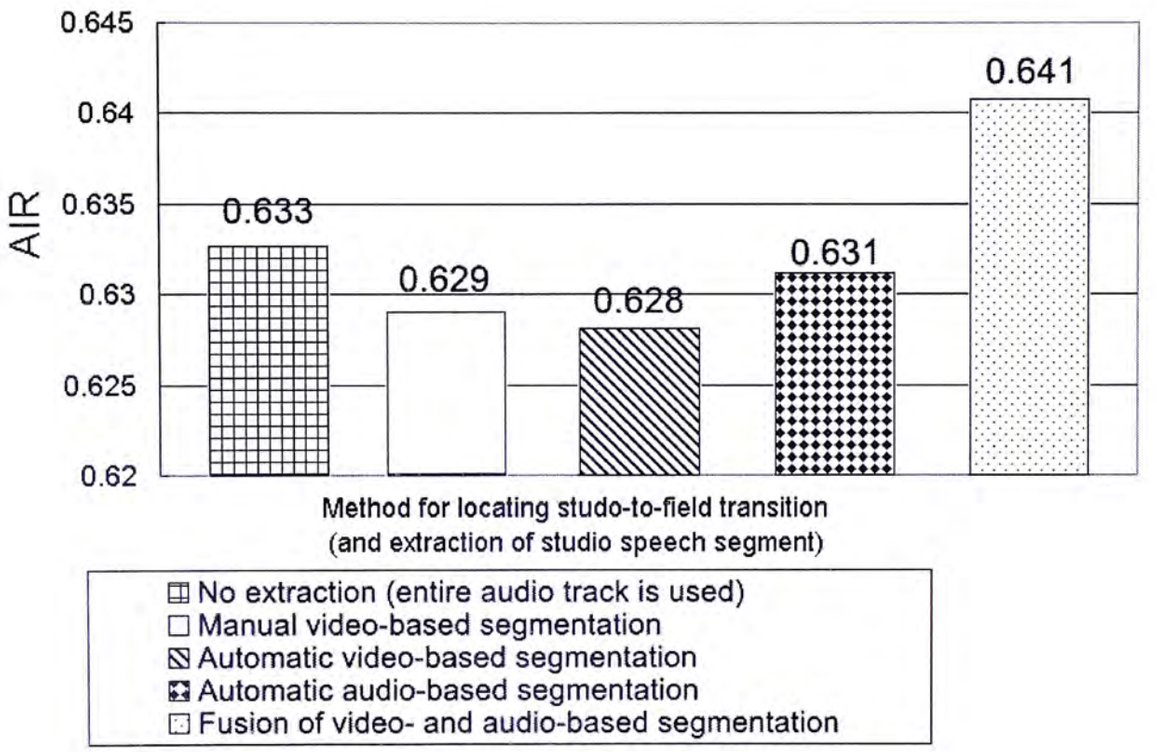

Figure 4.16: Retrieval performance based on extracted anchor/studio speech segments. Fusion of video- and audio-based segmentation gives the best retrieval result.

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.[7]

4.6 Chapter Summary

In this chapter, we have described the video-based segmentation algorithm. The algorithm has been used for automatic story segmentation, in response to the replacement of hand-segmentation of the news stories. The algorithm aims to detect field-to-studio transitions (story boundaries) in a news program. Results indicate that the algorithm can achieve a precision of 0.933 and a recall of 0.864 in story boundary detection.

Besides, we have devised three automatic methods to extract anchor speech from the audio tracks so as to reduce the adverse effect of speech recognition errors on retrieval performance. We have reported the design of three methods: (i) video-based segmentation aims to distinguish between the more homogeneous studio shots from the more dynamic field shots; (ii) audio-based segmentation uses Gaussian Mixture Models to distinguish the cleaner studio recordings from the noisier 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 $AIR=0.641$.

[7]The improvement is tested as statistically significant as shown in Appendix C.2.

Chapter 5

Document Expansion for Monolingual Spoken Document Retrieval

In this chapter, we describe our attempt to apply document expansion techniques [70] as our second robust method to monolingual spoken document retrieval (SDR). Document expansion is a technique used to enhance document retrieval by adding new terms that are probably relevant to the user-specified query to documents. Figure 5.1 illustrates the idea of document expansion, which expand documents in the collection with additional terms from external source(s). The use of document expansion technique aims to enrich the document representations and reduce the adverse effect of speech recognition errors on retrieval performance. We have performed document expansion using (i) selected field speech, (ii) N-best recognition hypotheses [71], and (iii) combined use of selected field speech and N-best recognition hypotheses. Selected field speech is the field speech extracted using annotations from Multimedia Markup Language (MmML). N-best recognition hypotheses are the N most probable syllable sequences output from the Cantonese base sylla-

ble recognizer. MmML stores detailed information of the speech segments, including dialect, type of speech, start and end time indices, etc. We can parse the MmML annotations to extract the field speech segments labeled with DIALECT="Cantonese". This method can help to extract Cantonese speech segments for retrieval experiments. Since we have used a Cantonese recognizer to process the speech data, there are many speech recognition errors from the non-Cantonese and non-speech segments in the field speech. The extraction of Cantonese speech segments may help to minimize the adverse effect of the speech recognition errors to 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. The transcribed queries may contain errors from homographs (i.e. a single character with multiple Cantonese pronunciations). The spoken documents are transformed into a syllable-based representation via speech recognition. The transcribed documents contain speech recognition errors. Hence the document syllables contain more transcription errors than query syllables. Document expansion using N-best recognition hypotheses may help to bridge this gap between queries and documents. Different weighting schemes have also been applied on the retrieval units so as to reflect their importance in document collection.

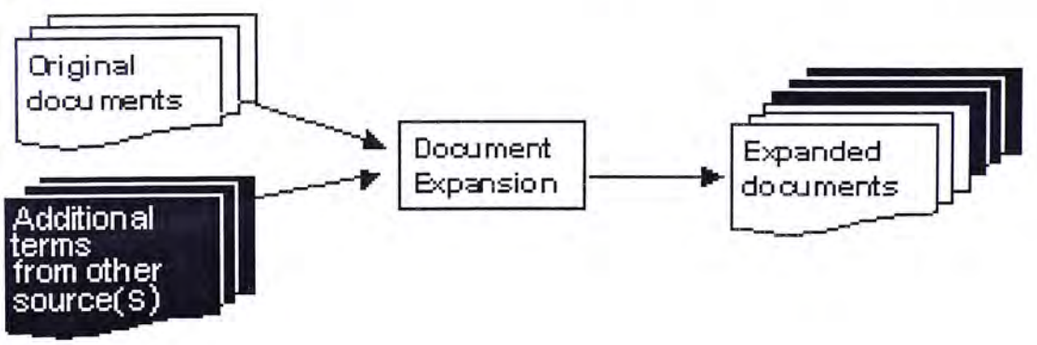

Figure 5.1: An illustration of the idea of document expansion. The original documents are expanded with additional terms from other source(s).

5.1 Document Expansion using Selected Field Speech Segments

Selected field speech segments are the useful speech segments extracted from field speech. "Useful" means the information extracted is beneficial to the retrieval task. While duration of an anchor speech segment constitutes approximately one fourth of a news story, field speech still contains much information about a news story. Document expansion using selected field speech segments aims to enrich the representations of anchor speech segments.

5.1.1 Annotations from MmML

In general, we can classify the field speech into the categories of reporter, interviewee, non-Cantonese and noise (non-speech segment) based on the annotations from MmML. Reporter and interviewee speech segments are presented in Cantonese by reporters and Cantonese-speaking interviewees respectively. With reference to Figure 5.2, the fourth layer of media component element is SpeakingStyle. The attribute TYPE is an indication of the property of the speech segment. TYPE may take on the values of ReporterSession (reporter speech), IntervieweeSession (interviewee speech) or NoiseSession (non-speech segment). The attribute DIALECT indicates the presentation language of the speech segment and may take on the values of "Cantonese", "English", "Mandarin", etc. After parsing, all the non-Cantonese speech segments are labeled as "foreign" as shown in Figure 5.3. The start and end time indices of these Cantonese speech segments are also provided by the MmML. We can extract the transcriptions of the speech segments by matching the time indices from MmML and recognition output.

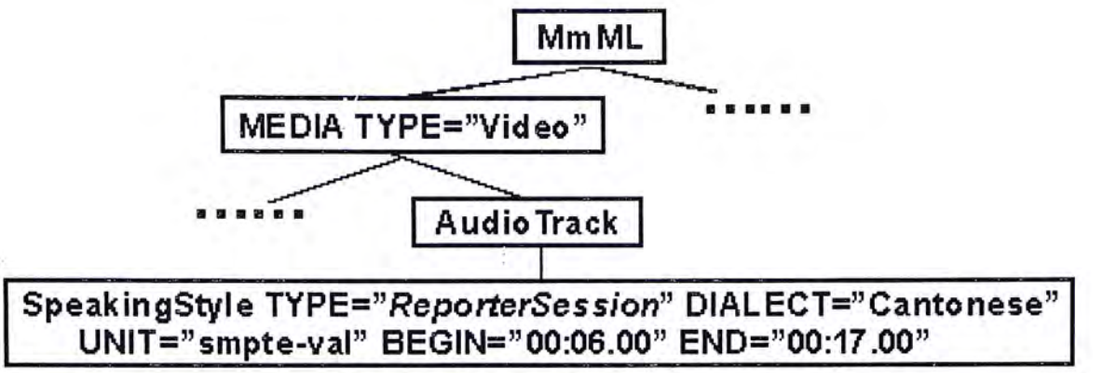

Figure 5.2: A simplified tree diagram showing that the element SpeakingStyle is at the fourth layer. SpeakingStyle contains the attributes TYPE and DIALECT to indicate the properties of speech segment.

Filename: 1999070701

.reporter

... lyun lung ngo jyu bou go ... (...聯龍娥茹報導...)

.foreign ⇐ non-Cantonese speech segment

... zyu cou gu haan bun gei gei ... (...鑄造故限搬幾幾...)

.interviewee

... gin ji dou bei fau kyut sei ... (...建議到被否決四...)

.noise

... saau haap deoi hyun fong cing ... (...稍狹對勸方稱...)

Figure 5.3: An example of the parsed output of different speech segments. Syllables after the segment labels (i.e. .reporter, .foreign, .interviewee and .noise) are recognized syllables that speech segments. The Chinese characters in the brackets are for reference and readability.

Combinations of speech segments	**AIR**
Anchor speech only (reference)	0.604
Anchor speech with reporter speech	**0.605**
Anchor speech with interviewee speech	0.572
Anchor speech with reporter and interviewee speech	0.603

Table 5.1: Spoken document retrieval performance based on different combinations of extracted field speech segments without re-weighting. The improvement is tested statistically significant using 0.3 level of signigicance in Figure C.1.

5.1.2 Selection of Cantonese Field Speech

Cantonese field speech includes reporter and interviewee speech. Due to the differences in acoustic conditions and speaking styles, the addition of retrieval terms from reporter and/or interviewee speech may harm the retrieval performance. We have performed a few experiments so as to verify the harmfulness of reporter and interviewee speech on retrieval performance. Three combinations of the speech segments have been used for retrieval experiments based on syllable bigrams. The results are shown in Table 5.1. The result without using any field speech segments is included as a reference (row 1 of Table 5.1). Results indicate that the use of reporter speech segments contain some useful information[8] (row 2 of Table 5.1). One possible reason is that the reporters have presented the details of the news in the field speech. Moreover, the speaking styles of the interviewees vary alot that increase the difficulty in speech recognition. Based on the observation, we perform the extraction of reporter speech for the following experiments.

[8]The improvement has been tested as statistically significant as shown in Appendix C.1.

5.1.3 Re-weighting Different Retrieval Units

We have performed re-weighting on the retrieval units according to their properties of recognition accuracy and duration. Anchor speech has higher syllable recognition accuracy than reporter speech (59.3% and 43.3% respectively). We have also found that the duration of selected field speech (reporter speech) is generally three times longer than that of anchor speech. We weighted the retrieval units from anchor speech ten times heavier than that from reporter speech in the document vector so as to reflect the importance of anchor speech. Ten is chosen because we have arbitrarily tested the values of one, five and ten and ten gave the best results among three. For simplicity, we used 10 and 1 as the weights of retrieval units from anchor speech and reporter speech respectively.

For retrieval purpose, every document is represented as a vector of syllable bigrams and skipped bigrams. We formed bigrams and skipped bigrams from the speech segments as shown in Figure 5.4 and re-weighted them as illustrated in Figure 5.5. Re-weighting of the different bigrams and skipped bigrams is based on the speech identity labeled (i.e. .anchor or .field). Since the retrieval units in the first two rows of Figure 5.5 are from anchor speech, they have the adjusted weights of ten. For the units from reporter speech, they have the weights of one.

5.1.4 Retrieval Performance with Document Expansion using Selected Field Speech

Additional terms will be added to the original documents during document expansion process. In this work, documents (anchor speech segments) are expanded according to the additional syllable bigrams and skipped bigrams derived from the reporter speech. The weighting function of the indexing terms (Equation 3.2) is modified as shown in Equation 5.1. The term frequency is

Filename: 1999070701

.anchor

... jik_wui wui_sei sei_nang jik_sei wui_nang ...

(...亦_會 會_四 四_能 亦_四 會_能...)

.reporter

... lyun_lung lung_ngo ngo_jyu (...聯_龍 龍_娥 娥_茹)

jyu_bou bou_go lyun_ngo (茹_報 報_告 聯_娥)

lung_jyu ngo_bou jyu_go ... (龍_茹 娥_報 茹_告...)

Figure 5.4: An example of the bigrams formed from the extracted output of reporter speech in Figure 5.3. The Chinese characters in the brackets are for readability only.

Filename: 1999070701

... jik_wui 10 wui_sei 10 (...亦_會 10 會_四 10)

sei_nang 10 jik_sei 10 wui_nang 10 (四_能 10 亦_四 10 會_能 10)

... lyun_lung 1 lung_ngo 1 ngo_jyu 1 (...聯_龍 1 龍_娥 1 娥_茹 1)

jyu_bou 1 bou_go 1 lyun_ngo 1 (茹_報 1 報_告 1 聯_娥 1)

lung_jyu 1 ngo_bou 1 jyu_go 1 (龍_茹 1 娥_報 1 茹_告 1)

Figure 5.5: An example of re-weighted document vector for experiments using anchor speech and reporter speech. The Chinese characters in the brackets are for readability. Since the syllable-character mapping is many-to-many, we are not able to present the *extact* content of the speech segments.

substituted with weighed term frequency.

$$d[i] = \ln(tw_d[i]) + 1.0 \tag{5.1}$$

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

We have performed automatic extraction of anchor speech segments using three different methods: (i) video-based segmentation, (ii) audio-based segmentation, and (iii) fusion of video- and audio-based segmentation (FVAS). Video-based segmentation algorithm utilizes information from the video tracks of the news data. Audio-based segmentation algorithm utilizes information from the audio tracks of the news data. FVAS algorithm combines the results from the algorithms (i) and (ii). The document expansion technique has been applied on the anchor speech segments from the automatic extraction methods.

Retrieval experiments have been performed using the expanded documents in the known-item retrieval (KIR) task described in Section 3.3. Recall that in the KIR task, only one document is considered to be relevant to a specific query. The retrieval experiment is based on the vector-space model. Average inverse rank (AIR) is used as the evaluation metric. Table 5.2 shows the retrieval performance with document expansion using selected field speech. Retrieval experiments are performed with the anchor speech extracted using various segmentation methods. The baseline result where no extraction method is used (i.e. entire audio track is used) and therefore without document expansion is included as a reference (row 1 of Table 5.2). The second column of Table 5.2 shows the retrieval with document expansion (labeled with "with"). The retrieval results without using document expansion are presented in the last column (labeled with "without").

Comparison between the second and third columns of Table 5.2 suggests that document expansion with reporter speech can improve the retrieval per-

Method for locating anchor-to-field transition (and extraction of anchor speech segment)	**AIR**	
	with	without
No extraction (entire audio track is used)	0.633	
Automatic video-based segmentation	0.679	0.628
Automatic audio-based segmentation	0.683	0.631
Fusion of video- and audio-based segmentation	**0.685**	0.641

Table 5.2: SDR performance based on extracted anchor speech segments with and without document expansion. Document expansion on anchor speech located using FVAS gives the best retrieval result. The improvement is tested as statistically significant at 0.01 level of significance.

formance. Document expansion with audio-based segmentation achieves slightly better performance than video-based segmentation. Document expansion with FVAS gives the best performance.[9]

5.2 Document Expansion using N-best Recognition Hypotheses

Our second attempt in document expansion is to expand the anchor speech segments using N-best recognition hypotheses. The N-best recognition hypotheses used for expansion are the N most probable syllable sequences output from the Cantonese base syllable recognizer. The indexing time required increases expotentially with the number of hypotheses generated by the recognizer. Therefore, we chose to use the five-best recognition hypotheses (i.e. $N = 5$) for expansion purposes. Figure 5.6 illustrates the expansion process using the five-best recognition hypotheses. All the documents in the collection have five-best recognition hypotheses. Differences among the N-best recogni-

[9]The improvement is tested as statistically significant as shown in Appendix C.3.

tion hypotheses may occur only in a few syllables. An example extracted from the N-best recognition output is shown in Figure 5.7.

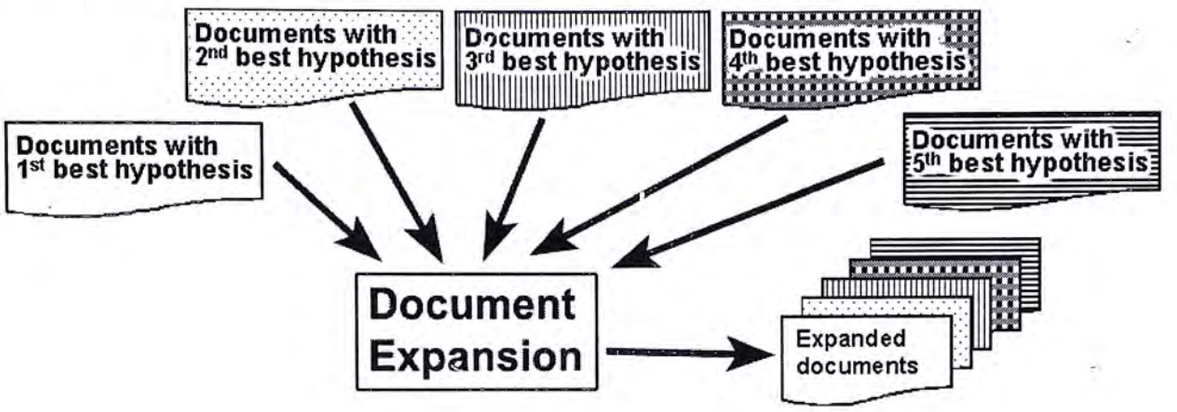

Figure 5.6: An illustration of document expansion using five-best recognition hypotheses. Expanded documents contain retrieval units from top-five recognition hypotheses.

Filename: 1999070701

1: ... jik wui **sei** nang ... (...亦 會 四 能...)

2: ... jik wui **sei** nang ... (...亦 會 四 能...)

3: ... jik wui **zau** nang ... (...亦 會 就 能...)

4: ... jik wui **sei** nang ... (...亦 會 四 能...)

5: ... jik wui **zau** nang ... (...亦 會 就 能...)

Figure 5.7: 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. The Chinese characters are for readability only.

We formed the bigrams and skipped bigrams from the N-best recognition hypotheses as shown in Figure 5.8.

Filename: 1999070701

1: ... jik_wui wui_**sei** **sei**_nang jik_**sei** wui_nang ...

(...亦_會 會_四 四_能 亦_四 會_能...)

2: ... jik_wui wui_**sei** **sei**_nang jik_**sei** wui_nang ...

(...亦_會 會_四 四_能 亦_四 會_能...)

3: ... jik_wui wui_**zau** **zau**_nang jik_**zau** wui_nang ...

(...亦_會 會_就 就_能 亦_就 會_能...)

4: ... jik_wui wui_**sei** **sei**_nang jik_**sei** wui_nang ...

(...亦_會 會_四 四_能 亦_四 會_能...)

5: ... jik_wui wui_**zau** **zau**_nang jik_**zau** wui_nang ...

(...亦_會 會_就 就_能 亦_就 會_能...)

Figure 5.8: An example on bigrams and skipped bigrams formed with the hypothesized syllables listed in Figure 5.7. The Chinese characters in the brackets are for readability.

5.2.1 Re-weighting Different Retrieval Units

Syllables that appear consistently across the N-best recognition outputs are likely to be more reliable and hence should be weighted more heavily in the document vector. The retrieval units in Figure 5.8 are re-weighted as illustrated in Figure 5.9. We used 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 retrieval unit is five. As shown in Figure 5.9, /wui_sei/ and /wui_zau/ have occurrences of three and two respectively in the hypotheses. Hence 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.

Filename: 1999070701

...jik_wui 5 wui_**sei** 3 wui_**zau** 2 **sei**_nang 3

(...亦_會 5 會_四 3 會_就 2 四_能 3)

zau_nang 2 jik_**sei** 3 jik_**zau** 2 wui_nang 5...

(就_能 2 亦_四 3 亦_就 2 會_能 5...)

Figure 5.9: An example on the re-weighting of different bigrams based on alternative recognition hypotheses in Figure 5.8. The Chinese characters are for readability.

5.2.2 Retrieval Performance with Document Expansion using N-best Recognition Hypotheses

The document expansion technique has been applied to the anchor speech segments from all of the three automatic extraction methods. Retrieval experiments have been performed using the expanded documents in the KIR task. Table 5.3 shows the retrieval results with and without document expansion

using N-best recognition hypotheses. The retrieval result with the entire audio track is included as a reference (row 1 of Table 5.3). The second column of Table 5.3 shows the retrieval with document expansion (labeled with "with"). The retrieval results without using document expansion are presented in the last column (labeled with "without").

Method for locating anchor-to-field transition (and extraction of anchor speech segment)	**AIR**	
	with	without
No extraction (entire audio track is used)	0.652	0.633
Manual video-based segmentation	0.639	0.629
Automatic video-based segmentation	0.639	0.628
Automatic audio-based segmentation	0.650	0.631
Fusion of video- and audio-based segmentation	**0.654**	0.641

Table 5.3: SDR performance based on extracted anchor speech segments with and without document expansion. Fusion of video- and audio-based segmentation gives the best retrieval result. The improvement is tested as statistically significant at 0.05 level of significance.

Comparison between the second and third columns of Table 5.3 suggests that the use of N-best recognition hypotheses for document expansion can consistently improve retrieval performance over the baseline (i.e. without using document expansion). Document expansion with audio-based segmentation achieves slightly better performance than video-based segmentation. A possible reason is that the$N$-best recognition hypotheses are from anchor speech segments only. The use of only anchor speech segments in retrieval may not necessarily improve retrieval performance over the reference performance (i.e. when the entire audio track is used). This is because 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. Document expansion with FVAS gives the best performance.[10]

5.3 Document Expansion using Selected Field Speech and N-best Recognition Hypotheses

Our third attempt in document expansion is to fuse information from selected field speech (reporter speech) and N-best recognition hypotheses. This work aims to take advantage from both of the information so as to further improve the retrieval performance.

5.3.1 Re-weighting Different Retrieval Units

We increase the weights of the N-best recognition hypotheses of anchor speech with respect to the number of occurrences of the token for retrieval. It is due to the reason that anchor speech has higher speech recognition accuracy than field speech. The weight of the retrieval units from reporter speech is increased to two. Figure 5.10 shows an example on the re-weighting of retrieval units. Since the units in first two rows are labeled as anchor speech with different occurrences, they have adjusted weights according to their occurrences as shown in Figure 5.9. For example, /wui_sei/ and /wui_zau/ have occurrences of three and two respectively in the hypotheses. Hence they have the adjusted weights of thirty and twenty. 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 the weights of fifty. Units from reporter speech are weighted as two.

[10]The improvement is tested as statistically significant as shown in Appendix C.4.

Filename: 1999070701

... jik_wui 50 wui_**sei** 30 wui_**zau** 20 **sei**_nang 30

(...亦_會 50 會_四 30 會_就 20 四_能 30)

zau_nang 20 jik_**sei** 30 jik_**zau** 20 wui_nang 50

(就_能 20 亦_四 30 亦_就 20 會_能 50...)

... lyun_lung 2 lung_ngo 2 ngo_jyu 2 jyu_bou 2

(...聯_龍 2 龍_娥 2 娥_茹 2 茹_報 2)

bou_go 2 lyun_ngo 2 lung_jyu 2 ngo_bou 2 jyu_go 2

(報_告 2 聯_娥 2 龍_茹 2 娥_報 2 茹_告 2)

Figure 5.10: Re-weighting the different bigrams and skipped bigrams based on alternative speech identity labeled and occurrences. The Chinese characters in side the brackets are for readability.

5.3.2 Retrieval Performance with Different Indexed Units

The document expansion has been applied to anchor speech segments extracted from the three automatic extraction methods. Table 5.4 shows the retrieval performance with document expansion on all of the three extraction methods. The performance of retrieval using entire audio track is included as a reference (row 1 of Table 5.4). The retrieval results without using any expansion method are included as the baseline (the last column, labeled with "without"). The second column of Table 5.4 (labeled with "combined") are the retrieval results with document expansion using both selected field speech and N-best recognition hypotheses. The third column (labeled with "field") are the retrieval performance with document expansion using selected field speech only. The fourth column (labeled with "N-best") are the retrieval performance with document expansion using N-best recognition hypotheses.

Comparison among columns two, three and four suggests that the use of

Method for locating anchor-to-field transition (and extraction of anchor speech segment)	**AIR**			
	combined	field	N-best	without
No extraction (entire audio track is used)	0.652	0.633	0.652	0.633
Automatic video-based segmentation	0.691	0.679	0.639	0.628
Automatic audio-based segmentation	0.691	0.683	0.650	0.631
Fusion of video- and audio-based segmentation	**0.694**	0.685	0.654	0.641

Table 5.4: Spoken document retrieval performance based on extracted anchor speech segments and different indexing terms. Document expansion with FVAS gives the best retrieval results. The improvement is tested as statistically significant at 0.01 level of significance.

selected field speech and N-best recognition hypotheses for document expansion can further improve the retrieval performance. Document expansion with audio-based segmentation achieves slightly better performance than video-based segmentation. Document expansion with FVAS gives the best performance.[11]

5.4 Chapter Summary

In this chapter, we attempt to improve retrieval performance using document expansion. Document expansion is a technique used to enhance document retrieval by adding potentially relevant terms to the documents. Document expansion is used to enrich document representation and may reduce the adverse effect of speech recognition errors on retrieval performance. We have performed document expansion using (i) selected Cantonese field speech, (ii)

[11]The improvement is tested as statistically significant as shown in Appendix C.5.

N-best recognition hypotheses, and (iii) the combination of (i) and (ii).

Document expansion using selected field speech segments aims to enrich the representations of anchor speech segments. Duration of anchor speech constitutes only one fourth of the entire news story. Field speech still contains much information about a news story. Selected field speech is the speech segments extracted from field speech that may be beneficial to the retrieval performance. In general, we can classify field speech into reporter, interviewee, non-Cantonese and non-speech segments. Reporter and interviewee speech are in Cantonese. Since we are focused on Cantonese spoken document retrieval, we only consider the speech segments from reporter and interviewee speech. We have performed extraction of field speech based on the annotations from Multimedia Markup Language (MmML) on our speech corpus. MmML is a convention of markup used to annotate the multimedia corpus we have collected. We can extract the Cantonese field speech segments with the annotations. We have chosen to use reporter speech for document expansion. This is because reporter speech has higher recognition accuracy than interviewee speech (43.3% vs 27%). We have also re-weighted the retrieval units based on their properties (anchor speech vs field speech). We have applied document expansion on three sources of anchor speech – anchor speech segments that are automatic extracted using (i) video-based segmentation, (ii) audio-based segmentation, and (iii) fusion of video- and audio-based segmentation (FVAS). Retrieval using expanded documents can consistently improve retrieval performance over the baseline. Results show that document expansion with FVAS gives the best performance. Document expansion using selected field speech segments can bring 8% improvement, achieving $AIR = 0.685$.

Our second attempt in document expansion used alternative recognition hypotheses to introduce additional indexing terms (syllable bigrams and skipped bigrams) to the extracted anchor speech. Retrieval experiments performed us-

ing five-best recognition hypotheses of anchor speech only. Results show that augmenting the top-scoring recognition hypotheses with N-best hypotheses brought consistent improvements over the baseline, with $AIR = 0.654$.

The combined use of selected field speech segments and N-best recognition hypotheses aims to improve retrieval performance further. Results show that the combined use can bring around 10% improvements over the baseline having $AIR = 0.694$.

Chapter 6

Query Expansion for Cross-language Spoken Document Retrieval

This chapter extends the study of query expansion [48] in monolingual spoken document retrieval (SDR) to cross-language SDR (CLSDR). Query expansion is a process adding potentially relevant new terms to a user-specified query. The intention is to to improve precision and/or recall by narrowing the lexical difference between queries and documents. The additional terms may be taken from a thesaurus, a side collection or specified relevant documents. The extra terms can have positive or negative weights. Positive weights mean the terms are found to be relevant and their weights in the query have been increased. Negative weights refer to the terms that are found to be irrelevant and their weights in the query have been decreased. The work in [48] demonstrated that query expansion is beneficial to monolingual SDR. This work extends the query expansion task from monolingual to cross-language. CLSDR is a retrieval task where queries and documents are in different languages. An illustration of a CLSDR task is given in Figure 6.1. In this work, English news stories (textual

queries) are used to retrieve Mandarin news broadcast (audio documents) in the archive. Prior to retrieval, English news stories are translated into Chinese textual queries. Thereafter, we have a monolingual Chinese SDR task.

Within-language ambiguity (WLA) and between language ambuguity (BLA) [72] are two common word mismatch problems in CLSDR. WLA is related to the meanings of word,[12] phrase, or sentence structure.[13] BLA is related to the expression of word, phrase or sentence in different languages. For example, a word can be translated in different ways with different meanings.[14] A phrase can be translated as a phrase or a series of individual words. In this work, the translation (mapping of words) is not a simple one-to-one mapping. Even if it is a one-to-one mapping, the words in the translated queries and the related documents may be lexicalized in different ways.

Much research effort has been devoted to disambiguation. Dictionary and corpus approaches are two common approaches. In dictionary approach, relationships between individual words are identified. This can be done by means of lexical information and structure analysis [73] [74]. In corpus approach, aligned and unaligned corpora are used for the analysis. A set of related documents, parallel corpora and comparable corpora are examples of aligned corpora. They are commonly used for the analysis of the most likely translations of terms and concepts between languages in the corpora [75]. There has also been work in the use of the co-occurrence method on unaligned corpora for term translation and target word selection [76] [77].

The queries are translated by referring to a English-Chinese bilingual terms

[12] A word may have multiple meanings. For example, *plane* can be either a noun or a verb. *Plane* (as a noun) can be *a flat surface, a level of development, an airplane, an airfoil*, etc. *Plane* (as a verb) can have the meaning of *to rise partly out of the water, to glide* or *to travel by airplane.*

[13] A sentence may lead to more than one interpretations. For example, "heating gas is dangerous" may be interpreted as *applying heat to gas is dangerous* or *gas used for heating is dangerous.*

[14] The word *organization* can be translated into 組織方法 (meaning: the act of organizing) or 團體 (meaning: an association).

list. The translated queries are from the Mandarin-English Information (MEI) project [30]. The documents are transcribed by means of Mandarin speech recognition. Transcriptions output by the recognition system are in simplified Chinese characters in GB coding. The bilingual terms list and Mandarin news data are from different sources. Therefore, there may be lexical differences between the translation dictionary and transcribed documents. Hence we try to explore the use of an automatic query expansion technique – pseudo-relevace feedback (PRF) to this CLSDR task. PRF can narrow the gap between keywords in the user-specified queries and document collection by expanding the query with terms from relevant documents in the collection. We have used a subset of TDT-2 corpus [78] as our experimental corpus in CLSDR. The details of the TDT-2 corpus will also be described.

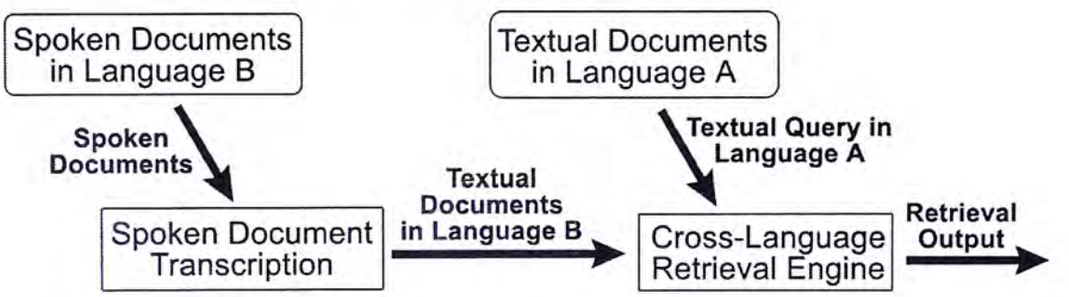

Figure 6.1: A general picture of a CLSDR task. Queries and documents are in different languages.

6.1 The TDT-2 Corpus

The corpus used for CLSDR task is the Topic Detection and Tracking Phase 2 (TDT-2) corpus. The materials in TDT-2 corpus include news articles and radio broadcasts from different news sources. In this task, we use a subset of the collection for the experiments: articles from New York Times Newswire Service (NYT) [79] and Associated Press Worldstream Service (APW) are

used as English textual queries for retrieval, and news recordings from Voice of America (VOA) [80] are used as Mandarin spoken documents. All the data were collected cover the six months period, from January 1998 to June 1998. Table 6.1 shows the details of the experimental corpus.

Source	**Collection period**	**Number of stories**
English textual queries		
New York Times and Associated Press	January 1998 to June 1998	195 stories (cover 17 topics)
Mandarin spoken documents		
Voice of America	March 1998 to June 1998	2,265 stories (cover 17 topics)

Table 6.1: Detailed information of the TDT-2 corpus used in the CLSDR experiments.

6.1.1 English Textual Queries

News articles from NYT and APW are used as textual queries in this CLSDR task. All of the articles are collected in the period from 4 January 1998 to 30 June 1998. The 195 articles are labeled as relevant to seventeen pre-selected topics[15] with the level of relevance.

There are at most twelve documents from each of the topic (see Appendix D). For the retrieval of Chinese documents, these queries need to be translated. This work uses the translated queries from the MEI project [30]. The named entities[16] in all the queries are tagged by the BBN Identifinder™[81]. The outputs are then sent to the phrase-based and word-based translation

[15]The list of topics covered are shown in Appendix D.

[16]Named entities include name expressions, time expressions and numeric expressions.

processes.[17] Multiple translation alternatives are also included. For retrieval task, the queries are divided into twelve batches, where each batch contains sixteen or seventeen documents from different topics.

6.1.2 Mandarin Spoken Documents

The spoken documents are from the VOA radio broadcast in TDT-2 collection. The news broadcasts are collected in the period from March 1998 to June 1998. The documents are indexed by a Chinese large-vocabulary continuous speech recognition (LVCSR) system from Dragon [82]. Transcriptions output by Dragon are segmented word sequences, which contain word boundaries and recognition errors.[18] There are 2,265 stories in the collection and are classified into the seventeen pre-selected topics as in the English textual queries.

6.2 Query Processing

Translated English queries from MEI project are adopted in the CLSDR experiments. An example on the Chinese translation of the query "counsel investigation of president clinton" is shown in Table 6.2.

6.2.1 Query Weighting

All translation alternatives are included in the retrieval task. Since the query vector considers the occurrences of terms in a query, the weight will be *unfair* among words with different number of translation alternatives. For example, in Table 6.2, the word "president" has twelve translation alternatives (the third row) while "clinton" only has one translation output (the last row).

[17] For example, the phrase "New Yorker" is translated as a phrase to 纽約客 in phrase-based translation process. However, the phrase is translated as "new" and "yorker" to 新的 and 貼面球 in word-based translated process.

[18] An example of speech recognition output is shown in Appendix E.

English word	**Chinese translation**	***weight***
counsel	参 法律顾问 商议 辩护人 劝告 忠告 劝导	$\frac{1}{7}$
investigation	调查 考察	$\frac{1}{2}$
president	主席 总统 议会 会长 总经理 董事长 大总统 总会会长 校长 知事 长官 总裁	$\frac{1}{12}$
clinton	克林顿	1

Table 6.2: An example on the multiple translation alternatives and weight adjustment on each of the Chinese translation alternatives.

Therefore, "president" will be weighted much more heavily than "clinton" in the vector-space model (VSM).

In order to minimize the ambiguity introduced by the translation alternatives, every translation is weighed inversely proportional to the number of translation alternatives.

$$weight_{translation} = \frac{1}{number\ of\ translation\ alternatives} \tag{6.1}$$

An example on the weight adjustment is shown in the last column of Table 6.2. The sum of the weights of each English word should be equal to one.

6.2.2 Bigram Formation

For the subword scale indexing, the translation alternatives are further expanded to form overlapping character bigrams. These subword units only form within the word boundaries, as shown in Table 6.3. Some of the translation alternatives will be ignored in the bigram formation process. A possible case is the present of a single character (see the double-underlined character in row 2 of Table 6.3). Underlined bigrams in row 3 are formed from the underlined translation alternative in row 2.

English word (**weight on word**)	counsel (1)
Chinese translation ($weight_{translation}$ **of each alternative**)	参 法律顾问 商议 辩护人 劝告 忠告 劝导 ($\frac{1}{7}$)
Bigrams ($weight_{bigram}$ **of every underlined bigram**)	法律 律顾 顾问 商议 辩护 护人 劝告 忠告 劝导 ($\frac{1}{7} \times \frac{1}{2} = \frac{1}{14}$)

Table 6.3: An illustration on the formation of bigrams from translation alternatives.

Each translation alternative contains a weight. After the formation of bigrams, the weight is further adjusted by dividing the $weight_{translation}$ of that translation alternative by the number of bigrams formed from it,[19] as shown in Equation 6.2.

$$\begin{aligned} weight_{bigram} &= \frac{1}{number\ of\ translation\ alternatives} \times \frac{1}{number\ of\ bigrams} \\ &= weight_{translation} \times \frac{1}{number\ of\ bigrams} \end{aligned} \tag{6.2}$$

6.3 Cross-language Retrieval Task

We use the TDT-2 collection to formulate a retrieval task. English textual queries are translated into Chinese, to retrieve Mandarin spoken documents. This work is also known as *query-by-example*, where English queries are used as an *exemplar* for the searching of documents in collection. Retrieved documents are in the same topic as the *exemplar*. Figure 6.2 illustrates the cross-language retrieval task in this work.

[19] Our previous work in [59] showed that weight adjustment can bring improvement in retrieval performance.

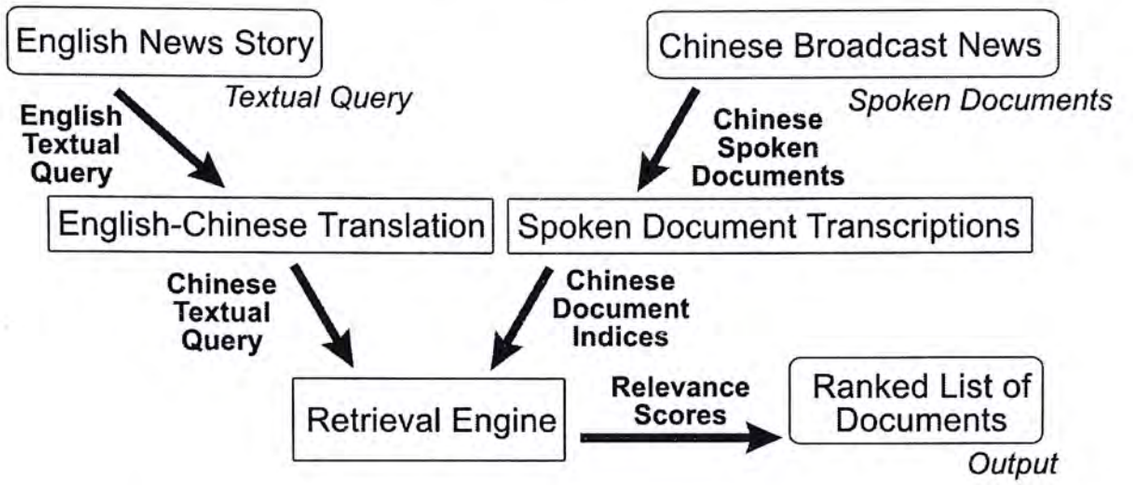

Figure 6.2: An illustration of the CLSDR task. An English news story is used to retrieve relevant Mandarin news broadcast.

6.3.1 Indexing Units

The use of overlapping syllable bigrams in indexing for retrieval can solve the Chinese word tokenization problem (refer to Section 1.2). The word boundaries of translated / transcribed documents are available in the TDT-2 corpus. We formulate the bigrams in the collections within a translated / transcribed word (see Table 6.3).

6.3.2 Retrieval Model

The retrieval model used in the experiments is also based on the vector-space model (VSM) as discussed in Section 3.4. In VSM, the similarity between query and document is computed as their inner product with cosine normalization (CN). However, in this work, we use document length normalization (DLN) instead of CN. This is because the previous work [59] has shown that DLN gave better retrieval performance on SDR over CN in a CLSDR task [59]. DLN is only based on the length of document and is more robust to recognition errors than CN [83]. The document vector (Equation 3.2) is modified as shown in Equation 6.3.

$$d[i] = \frac{1 + \ln tf_d[i]}{(1 - slope) \times length_{average} + slope \times length_{doc}} \tag{6.3}$$

where $length_{doc}$ is the length of the current document in term of bytes,

$length_{average}$ is the average document length across the collection, and

slope is used to control the proportion of distribution between $length_{doc}$ and $length_{average}$. *slope* can have a value between zero to one. For simplification, *slope* is set to 0.5 in this work.

6.3.3 Performance Measure

Every news article and broadcast in the TDT-2 corpus is annotated with a topic and the level of relevance. There are seventeen topics in total. The level of relevance can be either YES, BRIEF or irrelevant. YES for an article matched with a topic while BRIEF means the article is loosely related to that topic. A query-document pair is considered to be relevant if they are labeled as the same topic and both of them have the level of relevance YES. An example of the topic relevance table is shown in Table 6.4.

Since each query is relevant to more than one documents and relevance judgment is provided, mean average precision (mAP) is used as performance measure. We calculate the precision for every relevant document retrieved for a particular query and take an average of them to get the average precision value for that query. The average of the precision values obtained from all queries is the average precision for a particular topic. Taking another average over all topics produce a single value as the mAP, which is expressed as follows.

$$mAP = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{M_i} \sum_{j=1}^{M_i} \frac{1}{N_i} \sum_{k=1}^{N_i} \frac{k}{rank_{ijk}} \tag{6.4}$$

where L is the number of topic ($L = 17$ in this work),

M_i is the number of query for topic i,

N_i is the number of relevant documents for topic i, and

Topic relevance table

Query/Document Number	Topic ID	Level of Relevance
APW19980127.0599	1	YES
NYT19980104.0075	1	BRIEF
VOA19980104.2300.0904	1	YES
VOA19980105.2100.1684	1	BRIEF

Relevance information derived

	VOA19980104.2300.0904	VOA19980105.2100.1684
APW19980127.0599	relevant	irrelevant
NYT19980104.0075	irrelevant	irrelevant

Table 6.4: An example of the topic relevance table and the relevance information derived.

$rank_{ijk}$ is the rank of the k^{th} relevant document in the ranked list for query j on topic i.

6.4 Relevance Feedback

Th problem of word mismatch occurs when the keyword(s) chosen by the user in the query differ from the keyword(s) contained in the relevant document. This may lead to failure in retrieving the relevant document(s) needed by the user.

Query expansion is one of the techniques that can be used to solve the word mismatch problem. It can be done by relevance feedback, which expands the query with terms related to the document collection. Relevance feedback has been widely used in different areas, including image search [84], music retrieval [85], document filtering [86] and cross-language information retrieval [87]. The relevance feedback algorithm proposed by Rocchio is one of the commonly used

algorithms [88]. In Rocchio's algorithm, the initial query can be automatically adjusted as shown in Equation 6.5. The importance of query terms present in the relevant documents are increased by increasing the terms' weights or introducing some new terms to the original query. The importance of terms co-exist with the terms from the non-relevant documents are reduced. This can be achieved by decreasing the terms' weights or removing them from the original query.

$$q_{new} = \alpha q + \beta(\frac{1}{N_r} \sum_{i \in D_r} d_i) - \gamma(\frac{1}{N_n} \sum_{j \in D_n} d_j) \tag{6.5}$$

where q and q_{new} are the query vectors before and after query expansion,

D_r and N_r denote the set and the number of relevant documents,

D_n and N_n denote the set and the number of non-relevant documents, and

α, β and γ are tunable parameters that used to control the relative effects of the original, added and removed terms, respectively.

6.4.1 Pseudo-Relevance Feedback

In the PRF algorithm, the initial search is performed for the query. The top N_r retrieved documents are *assumed* to be relevant (therefore it is "pseudo") and the bottom N_n documents are *assumed* to be non-relevant. We can control the number of terms that are added to N_{rt} or removed from N_{nt} the original query. The query can be automatically adjusted using these relevant judgements. After the process, the query terms get re-weighted. An illustration of PRF is shown in Figure 6.3.

6.5 Retrieval Performance

For simplicity, α, β and γ in Equation 6.5 are set to one in the experiments. We have carried out a few experiments to find out the "optimal" values for

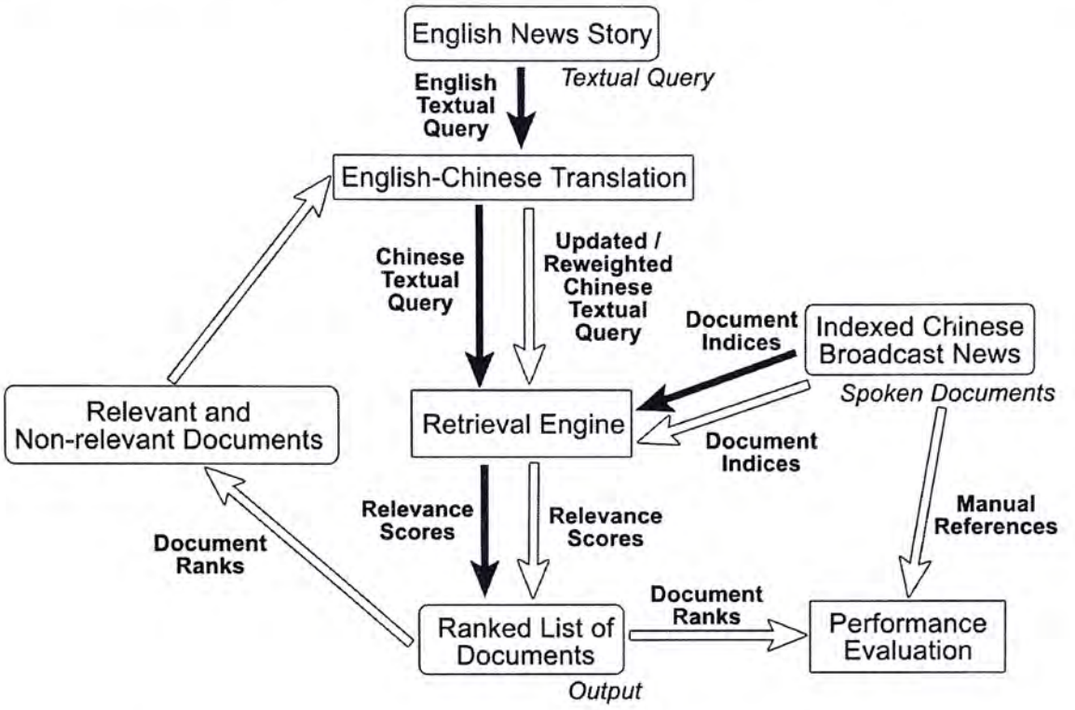

Figure 6.3: A illustration of pseudo-relevance feedback algorithm in CLSDR experiments.

N_r, N_{rt}, N_n and N_{nt}.[20] The optimal values for N_r, N_{rt}, N_n and N_{nt} have been found to be 2, 120, 1 and 50 respectively.

Retrieval performance with and without the use of PRF is given in Figure 6.4. Recall that the translated queries are divided into twelve batches for retrieval experiments. Results show that, for the experiment without query expansion with PRF, the values of the average precision (AP) fall in the range between 0.351 and 0.479. The average of all twelve query batches is $mAP = 0.410$. This is the baseline reference. For the experiments with PRF, the range of AP for the expanded queries range between 0.421 and 0.573. The average of the expanded queries is $mAP = 0.514$. The result indicates that the use of PRF can improve retrieval performance across all query batches.[21]

A detailed list of the numbers is shown in Table 6.5. Retrieval performance

[20]The preliminary results on the parameter estimation experiments are listed in Appendix F.

[21]The improvement has been tested as statistically significant as shown in Appendix C.6.

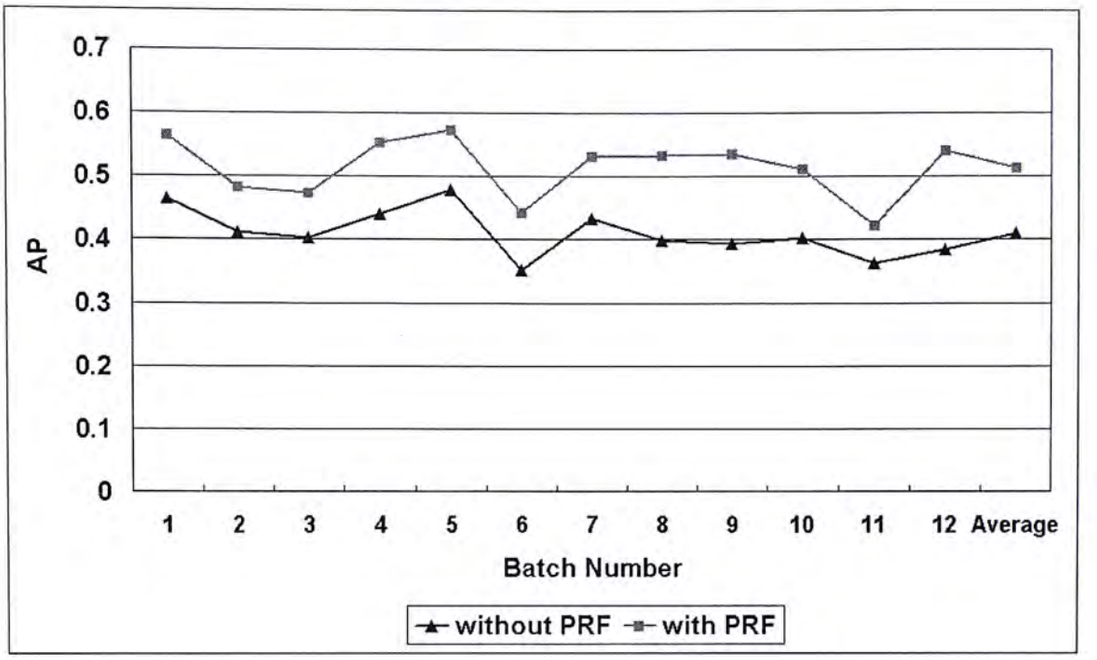

Figure 6.4: A comparison between baseline and query expansion with PRF across all query batches (in AP) and the average value (in mAP) in CLSDR task.

with PRF for query expansion can be improved by 25.1% . The improved retrieval performance is due to the introduction of "more related terms" with documents to query batches during expansion.

6.6 Chapter Summary

In this chapter, we have extended our study of query expansion from monolingual spoken document retrieval (SDR) to cross-languague spoken document retrieval (CLSDR). Query expansion is the process adding new terms to a user-specified queries. Previous work demonstrated that query expansion can bring improvement to monolingual SDR. CLSDR is a retrieval task where the queries and documents are in different languagues. Within-language ambiguity (WLA) and between-language ambiguity (BLA) are common word mismatch problems in CLSDR. Our CLSDR task uses English textual queries to retrieve Mandarin spoken documents. English queries are translated into

Batch Number	1	2	3	4	5	6	7	8
without PRF	0.464	0.411	0.403	0.441	0.479	0.351	0.433	0.398
with PRF	0.564	0.482	0.474	0.554	0.573	0.443	0.532	0.534

Batch Number	9	10	11	12	Average over 12 batches
without PRF	0.393	0.402	0.363	0.385	**0.410**
with PRF	0.535	0.511	0.422	0.540	**0.514**

Table 6.5: Retrieval performance for twelve query batches (in *AP*) and the average value (*mAP*) over all the batches. The improvement in *mAP* is tested as statistically significant at 0.01 level of significance.

Chinese prior to retrieval. All the translation alternatives are included in the translated queries and the queries are re-weighted. Mandarin documents are automatically transcribed with a Chinese large-vocabulary continous speech recognition (LVCSR) system. Both translated queries and documents are indexed with overlapping characters bigrams for retrieval. We have explored the use of an automatic query expansion method – pseudo-relevance feedback (PRF) in the CLSDR task.

PRF is used to address the WLA and BLA problems by expanding the queries with terms from the top-ranking documents. PRF can add some potentially *relevant* terms to the original query and remove some potentially *non-relevant* terms from the query. Results show that the retrieval performance of CLSDR task has a baseline of *mAP=0.410*. Query expansion using PRF can bring 25.1% improvement when compared with the baseline and achieving *mAP=0.514*.

Chapter 7

Conclusions and Future Work

This thesis explored different robust techniques for Chinese spoken document retrieval (SDR). SDR is the task of automatically retrieving relevant spoken documents with respect to user-specified requests. The key components in a SDR system include automatic speech recognition (ASR) and information retrieval (IR). ASR is used to extract information from the spoken documents and represent them as textual information. IR technique is used to retrieve relevant information base on the queries. Robust methods proposed in this work are the methods that can endure the adverse factors from ASR and IR, and maintain (or even enhance) the SDR performance.

This work is motivated by the improved retrieval performance achieved from fusion of different sources of information. We have studied three robust methods in this work: (i) automatic extraction of anchor speech, (ii) document expansion for monolingual SDR, and (iii) query expansion for cross-language SDR (CLSDR). Automatic extraction of anchor speech is achieved by the fusion of information from audio and video tracks. Speech recognition accuracy affects retrieval performance. The speech recognition accuracy of anchor speech is higher than other speech segments. The extraction of anchor speech for IR should be beneficial to monolingual SDR performance. We

have investigated three attempts in automatic extraction of anchor speech: (i) video-based, (ii) audio-based, and (iii) fusion of video and audio. Video-based segmentation utilizes the video frame information of the news data. The algorithm aims to distinguish the more homogeneous anchor shots from the more dynamic field shots. Audio-based segmentation performs speech classification using audio track information. The classification process uses Gaussian Mixture Models to distinguish the cleaner anchor speech (in studio-quality) from the noisier field speech. A fusion strategy combines video- with audio-based segmentation to achieve precise extraction of anchor speech. The total duration of anchor speech is only one fourth of the entire collection. Indexing of only anchor speech can reduce the indexing effort required and the amount of data involved by a factor of four.

The second robust method – document expansion is achieved by the fusion of information from selected field speech and N-best recognition hypotheses. Document expansion is a technique used to enhance document retrieval by adding more relevant terms to documents. We have explored three attempts in document expansion: (i) the use of selected field speech, (ii) the use of N-best recognition hypotheses, and (iii) combination of both. Duration of anchor speech is only one fourth of the entire collection. Field speech still contains much information about a news story. We have annotated the information and properties of the field speech using Multimedia Markup Language (MmML). MmML is a set of convention, which is used to annotate the news data collection. MmML stores the dialect, type of speech, start and end time indices of the speech segments. In general, field speech can be classified into reporter, interviewee, non-Cantonere and non-speech segments. Reporter speech has the highest recognition accuracy among them. We have also found that reporter speech segments contain information that is beneficial to retrieval performance. Therefore, we have performed extraction of reporter

speech using annotations from MmML. Document expansion using selected field speech segments can reduce the adverse effect of noisy speech to retrieval performance. N-best recognition hypotheses are the N most probable syllable sequences output from our Cantonese base syllable recognizer. Document expansion using N-best recognition hypotheses can reflect the reliability of the recognition hypotheses. Document expansion using both of the information can take advantage of Cantonese field speech segments and N-best recognition hypotheses.

Query expansion is a technique used to narrow the lexical difference between queries and documents. Query expansion is achieved by the fusion of information of the ranked list of retrieved documents in initial search. The work in [48] showed that query expansion can bring improvement to monolingual SDR. The work extends the work from monolingual to cross-language. CLSDR is a retrieval task where the queries and documents are in different languages. In our work, the English queries are translated into Chinese for the retrieval of Mandarin spoken documents. Within-language ambiguity (WLA) and between-language ambiguity (BLA) are two common word mismatch problems in CLSDR. Query expansion can reduce the lexical difference between queries and documents by adding relevant terms to queries. One technique for automatic query expansion is pseudo-relevance feedback (PRF). PRF is used to address the WLA and BLA problems by expanding the queries with terms from top-ranking documents. PRF can automatically refine the queries by promoting the relevant terms and demoting the irrelevant terms in the queries.

In summary, we have made the following contributions:

- Design of MmML to promote the annotation of multilingual multimedia information in a structural and meaningful way. MmML is a markup language, which is designed with reference to SMIL 2.0 specifications

and XML schema. MmML is able to manage and store the multilingual multimedia information.

- Fusion of video- and audio-based segmentation algorithm is proven to be favorable to monolingual SDR. The robust fusion method can extract the reliable speech data (anchor speech) precisely. It is found that with the precise extration of speech data, the use of only anchor speech can improve the retrieval performance.

- Document expansion can further improve retrieval performance for monolingual SDR. Document expansion is the second robust method proposed in this work. Document expansion is performed with reporter speech and N-best recognition hypotheses. Anchor speech is expanded to include the top-five recognition hypotheses. This is further expanded with reference to MmML annotations. Different weighting schemes have been used to appropriately reflect the importance of retrieval units from anchor speech.

- The third robust method described is automatic query expansion using PRF. PRF is testified to be beneficial to CLSDR task. PRF is used to reduce the lexical difference between the translated queries and transcribed documents. PRF expands the user-specified query with the terms from the documents collection based on a ranked list of retrieved documents. PRF also removes some potentially irrelevant terms from the original query.

7.1 Future Work

In the current monolingual SDR task, we used the summary title as the query to retrieve its corresponding spoken document from the collection. The summary title is only a short sentence, which contains limited number of keywords and may contain many Chinese abbreviations. A possible direction in this area

is to apply query expansion using side collection so as to enrich the representation of and the keywords contained in the query.

The use of SDR technique as document classification is another extension of the current research. We can use a query to retrieve its best-matched document and use the relevant document to find other news broadcasts that are similar to it. A threshold can be used to control the degree of similarity between news broadcasts or the number of nearest neighbors around the relevant document. Therefore, we can group the news broadcasts into topics with certain similarity.

A possible area for further work in CLSDR is to investigate the variations of the term weighting scheme. Our current weighting scheme set the weightings of the original α, relevant β and non-relevant terms γ to one. It may not be a good way to weight the terms in the initial query as important as the newly added / removed ones. In addition, emphasis should be put on the relevant terms when compare with the non-relevant terms. The effect on the negative feedback can be further studied.

Another suggestion is to increase the number of feedback iterations in PRF. New terms from the top-ranking documents are added to the original query in each iteration. There should be a optimal number of iterations in relevance feedback for the addition of new terms. The optimal number mat be related to the length of the original query. With reference to the work in [89], we expect four or five will be a preferrable number of iteration.

In current work, we only use the top-N_r[22] ranking documents as relevant documents and in equal weighting in relevance feedback. We can try to use the ranked output from the initial search as a relevance scale. We can apply different weightings on the terms from documents in different ranks. The retrieval list is ranked according to the degree of similarity between query and documents. Therefore, the terms from document at rank one should be

[22] N_r is the number of top-ranking documents assumed to be relevant in the PRF experiments.

weighted more heavily when compare with the terms from document at other ranks.

Another possible direction is to solve the problem of ambiguity in the translation process. Currently, our algorithm includes all translation alternatives and apply different weightings on the translated terms. We can use a side collection as parallel corpus for the co-occurrence frequency measuement. The co-occurrence frequency method can help to find out the target selection term for retrieval.

Appendix A

XML Schema for Multimedia Markup Language

Figure A.1 shows the XML schema associated with the design of Multimedia Markup Language (MmML). The XML schema shows the definitions and all the relationships between the elements and attributes in MmML. xs:element defines the element type and content. An element can contain child elements (no character data), string (labeled with xs:string) or integer (label with xs:integer) but no other elements. minOccurs and maxOccurs indicate the minimum and maximum number of the occurrences of that element / attribute. unbounded means there is no limitation on the occurrences. xs:attribute is used to declare the attributes associated name-value pairs with elements. required labels that the attribute must be always present while optional means optional. Figure A.2 is a diagram of the design of MmML.

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
 <xs:element name="MmML">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="MEDIA" maxOccurs="unbounded"/>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
 <xs:element name="MEDIA">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="TimeStamp" minOccurs="0" maxOccurs="unbounded"/>
    <xs:element ref="VideoTrack" minOccurs="0"/>
    <xs:element ref="AudioTrack" minOccurs="0"/>
    <xs:element ref="TextualInfo" minOccurs="0" maxOccurs="unbounded"/>
   </xs:sequence>
   <xs:attribute name="TYPE" type="xs:string" use="required"/>
   <xs:attribute name="LocalRef" type="xs:string" use="required"/>
   <xs:attribute name="StyleRef" type="xs:string" use="optional"/>
   <xs:attribute name="Style" type="xs:string" use="required"/>
  </xs:complexType>
 </xs:element>
 <xs:element name="TimeStamp">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="Date">
     <xs:simpleType>
```

```
        <xs:restriction base="xs:date">
          <xs:pattern value="\d{4}\d{2}-\d{2}"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:element>
    <xs:element ref="Time" minOccurs="0">
      <xs:simpleType>
        <xs:restriction base="xs:time">
          <xs:pattern value="\d{2}:\d{2}:\d{2}"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:element>
  </xs:sequence>
  <xs:attribute name="TYPE" type="xs:string" use="required"/>
 </xs:complexType>
</xs:element>
<xs:element name="VideoTrack">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="VCoding"/>
   <xs:element ref="Shot" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="UNIT" type="xs:string" use="required"/>
  <xs:attribute name="BEGIN" type="xs:string" use="required"/>
  <xs:attribute name="END" type="xs:string" use="required"/>
 </xs:complexType>
</xs:element>
<xs:element name="AudioTrack">
 <xs:complexType>
```

```
    <xs:sequence>
      <xs:element ref="ACoding"/>
      <xs:element ref="SpeakingStyle" maxOccurs="unbounded"/>
    </xs:sequence>
    <xs:attribute name="UNIT" type="xs:string" use="required"/>
    <xs:attribute name="BEGIN" type="xs:string" use="required"/>
    <xs:attribute name="END" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
<xs:element name="TextualInfo">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="TCoding" minOccurs="0"/>
      <xs:element ref="Title" minOccurs="0"/>
      <xs:element ref="Content" minOccurs="0" maxOccurs="unbounded"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="VCoding">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="VStream" minOccurs="0"/>
      <xs:element ref="FrameSize" minOccurs="0"/>
      <xs:element ref="AspectRatio" minOccurs="0"/>
      <xs:element ref="Standard" minOccurs="0"/>
      <xs:element ref="FrameRate" minOccurs="0"/>
      <xs:element ref="BitRate" minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
```

```
</xs:element>
<xs:element name="Shot">
 <xs:complexType>
  <xs:sequence minOccurs="0" maxOccurs="unbounded">
   <xs:element ref="Anchor" minOccurs="0" maxOccurs="unbounded" />
   <xs:element ref="Reporter" minOccurs="0" maxOccurs="unbounded" />
   <xs:element ref="Interviewee" minOccurs="0" maxOccurs="unbounded" />
   <xs:element ref="Empty" minOccurs="0" maxOccurs="unbounded" />
  </xs:sequence>
  <xs:attribute name="NAME" type="xs:string" use="required" />
  <xs:attribute name="UNIT" type="xs:string" use="required" />
  <xs:attribute name="BEGIN" type="xs:string" use="required" />
  <xs:attribute name="END" type="xs:string" use="required" />
 </xs:complexType>
</xs:element>
<xs:element name="ACoding">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="AStream" minOccurs="0" />
   <xs:element ref="Sampling" minOccurs="0" />
   <xs:element ref="Channel" minOccurs="0" />
   <xs:element ref="BitRate" minOccurs="0" />
  </xs:sequence>
 </xs:complexType>
</xs:element>
<xs:element name="SpeakingStyle">
 <xs:complexType>
  <xs:sequence minOccurs="0" maxOccurs="unbounded">
   <xs:element ref="Anchor" minOccurs="0" maxOccurs="unbounded" />
```

```
      <xs:element ref="Reporter" minOccurs="0" maxOccurs="unbounded" />
      <xs:element ref="Interviewee" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    <xs:attribute name="TYPE" type="xs:string" use="required" />
    <xs:attribute name="DIALECT" type="xs:string" use="optional" />
    <xs:attribute name="UNIT" type="xs:string" use="required" />
    <xs:attribute name="BEGIN" type="xs:string" use="required" />
    <xs:attribute name="END" type="xs:string" use="required" />
  </xs:complexType>
</xs:element>
<xs:element name="TCoding">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Encoding" />
      <xs:element ref="Platform" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="Title">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:string">
        <xs:attribute name="UNIT" type="xs:string" use="required" />
        <xs:attribute name="BEGIN" type="xs:integer" use="required" />
        <xs:attribute name="END" type="xs:integer" use="required" />
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
```

```
<xs:element name="Content">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="TYPE" type="xs:string" use="required" />
    <xs:attribute name="UNIT" type="xs:string" use="optional" />
    <xs:attribute name="BEGIN" type="xs:integer" use="optional" />
    <xs:attribute name="END" type="xs:integer" use="optional" />
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
<xs:element name="VStream" type="xs:string" default="MPEG-1" />
<xs:element name="FrameSize">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="UNIT" use="required" />
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
<xs:element name="AspectRatio" type="xs:string" default="4:3" />
<xs:element name="Standard" type="xs:string" default="PAL" />
<xs:element name="FrameRate">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:integer">
    <xs:attribute name="UNIT" type="xs:string" use="required" />
```

```
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="BitRate">
  <xs:complexType>
    <xs:simpleContent>
      <xs:extension base="xs:integer">
        <xs:attribute name="UNIT" type="xs:string" use="required"/>
      </xs:extension>
    </xs:simpleContent>
  </xs:complexType>
</xs:element>
<xs:element name="Anchor">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="PersonalDetails"/>
    </xs:sequence>
    <xs:attribute name="UNIT" type="xs:string" use="optional"/>
    <xs:attribute name="BEGIN" type="xs:string" use="optional"/>
    <xs:attribute name="END" type="xs:string" use="optional"/>
  </xs:complexType>
</xs:element>
<xs:element name="PersonalDetails">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Gender">
        <xs:simpleType>
          <xs:restriction base="xs:string">
```

```
            <xs:enumeration value="Female"/>
            <xs:enumeration value="Male"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:element>
      <xs:element name="Name"type="xs:string"minOccurs="0"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="Reporter">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="PersonalDetails"/>
    <xs:sequence>
    <xs:attribute name="UNIT"type="xs:string"use="optional"/>
    <xs:attribute name="BEGIN"type="xs:string"use="optional"/>
    <xs:attribute name="END"type="xs:string"use="optional"/>
  </xs:complexType>
</xs:element>
<xs:element name="Interviewee">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="PersonalDetails"/>
    </xs:sequence>
    <xs:attribute name="UNIT"type="xs:string"use="optional"/>
    <xs:attribute name="BEGIN"type="xs:string"use="optional"/>
    <xs:attribute name="END"type="xs:string"use="optional"/>
  </xs:complexType>
```

```
</xs:element>
<xs:element name="Empty">
 <xs:complexType>
  <xs:sequence>
   <xs:element name="SoundType"/>
  </xs:sequence>
  <xs:attribute name="UNIT" type="xs:string" use="optional"/>
  <xs:attribute name="BEGIN" type="xs:string" use="optional"/>
  <xs:attribute name="END" type="xs:string" use="optional"/>
 </xs:complexType>
</xs:element>
<xs:element name="AStream" type="xs:string" default="MPEG Audio
 Layer II"/>
<xs:element name="Sampling">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="UNIT" use="required"/>
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
<xs:element name="Channel" type="xs:string" default="Stereo"/>
<xs:element name="Encoding" type="xs:string" default="Unicode"/>
<xs:element name="Platform" type="xs:string" default="Windows"/>
</xs:schema>
```

Figure A.1: The XML schema – xml_schema.xsd, for MmML.

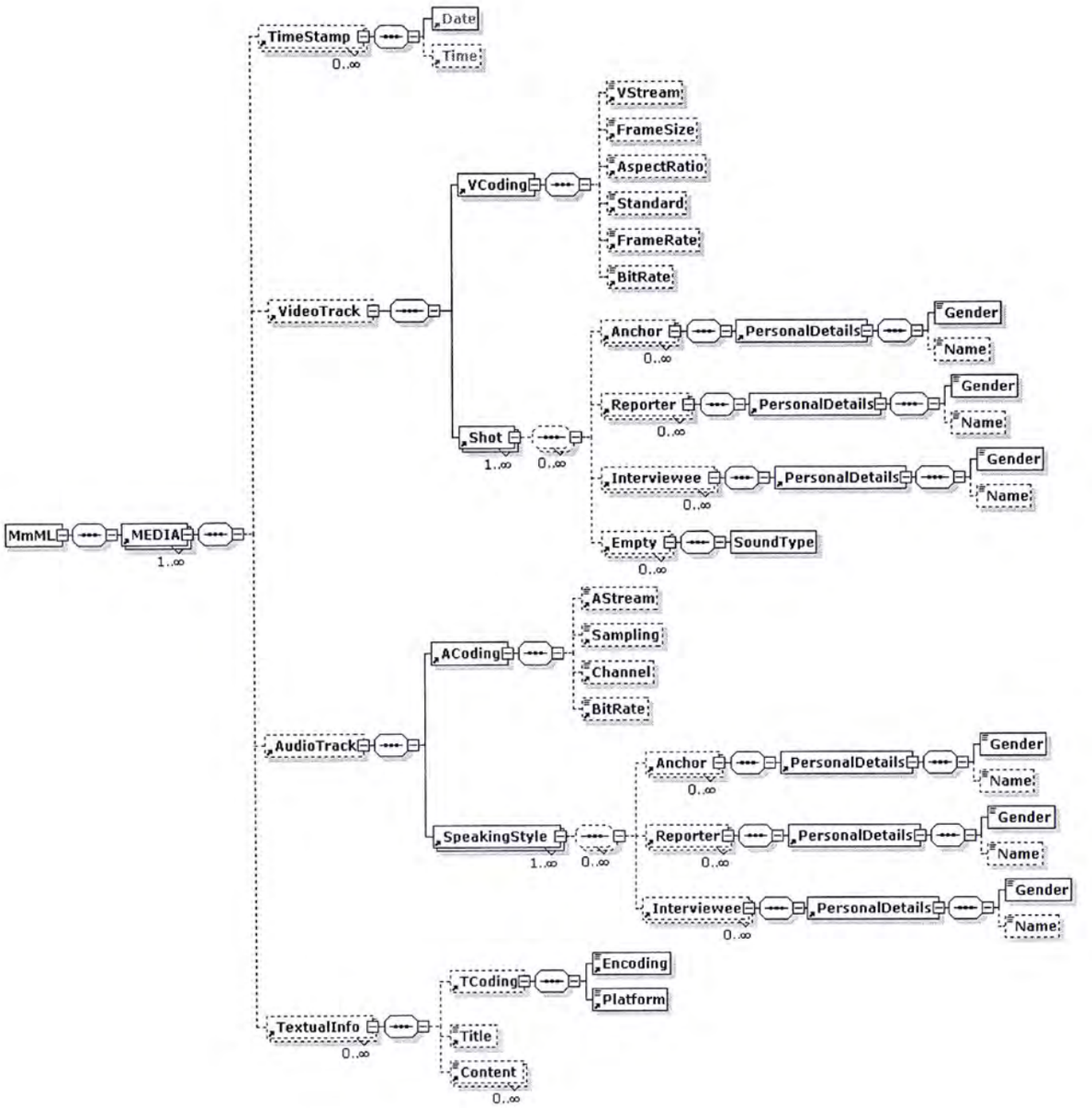

Figure A.2: An illustration of the full picture of MmML.

Appendix B

Example of Multimedia Markup Language

Figure B.1 is an example of MmML for a news story with filename 1999080409 (corresponding to the ninth story on August 4, 1999). The first layer (labeled with MEDIA) shows a video file with its corresponding text file. The video file contains both video and audio tracks. The text file contains the textual summary, title and copyright information of the news story.

```
<?xml version="1.0"standalone="yes "?>
<MmML xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance "
xsi:noNamespaceSchemaLocation= "L:\xml_ schema.xsd">
 <MEDIA TYPE="Video"LocalRef="1999080409.mpg"Style="NewsReport">
  <TimeStamp TYPE="SourceDate">
   <Date>1999-08-04</Date>
  </TimeStamp>
  <VideoTrack UNIT="smpte-val"BEGIN="00:00.00"END="02:01.00">
   <VCoding>
    <VStream>MPEG-1</VStream>
     <FrameSize UNIT="pixels">352x288</FrameSize>
     <AspectRatio>4:3</AspectRatio>
     <Standard>PAL</Standard>
     <FrameRate UNIT="fps">25</FrameRate>
     <BitRate UNIT="kbit/s">1150</BitRate>
    </VCoding>
    <Shot NAME="AnchorShot"UNIT="smpte-val"BEGIN="00:00.00"
    END="00:17.00">
     <Anchor UNIT="smpte-val"BEGIN="00:00.00"END="00:06.00">
      <PersonalDetails>
       <Gender>Female</Gender>
       <Name>鄧淑芳</Name>
      </PersonalDetails>
     </Anchor>
     <Anchor UNIT="smpte-val"BEGIN="00:06.00"END="00:17.00">
      <PersonalDetails>
       <Gender>Male</Gender>
       <Name>李燦榮</Name>
```

```
    </PersonalDetails>
   </Anchor>
  </Shot>
  <Shot NAME="FieldShot" UNIT="smpte-val" BEGIN="00:17.00"
  END="02:01.00">
   <Reporter UNIT="smpte-val" BEGIN="00:17.00" END="00:21.00">
    <PersonalDetails>
     <Gender>Female</Gender>
     <Name>莫宜端 </Name>
    </PersonalDetails>
   </Reporter>
   <Empty UNIT="smpte-val" BEGIN="00:21.00" END="00:24.00">
    <SoundType>Noise</SoundType>
   </Empty>
   <Reporter UNIT="smpte-val" BEGIN="00:24.00" END="00:47.00">
    <PersonalDetails>
     <Gender>Female</Gender>
     <Name>莫宜端 </Name>
    </PersonalDetails>
   </Reporter>
   <Interviewee UNIT="smpte-val" BEGIN="00:47.00" END="01:08.00">
    <PersonalDetails>
     <Gender>Male</Gender>
    </PersonalDetails>
   </Interviewee>
   <Reporter UNIT="smpte-val" BEGIN="01:08.00" END="01:40.00">
    <PersonalDetails>
     <Gender>Female</Gender>
     <Name>莫宜端 </Name>
```

```
    </PersonalDetails>
   </Reporter>
   <Interviewee UNIT="smpte-val" BEGIN="01:40.00" END="01:49.00">
    <PersonalDetails>
     <Gender>Male</Gender>
    </PersonalDetails>
   </Interviewee>
   <Reporter UNIT="smpte-val" BEGIN="01:49.00" END="02:01.00">
    <PersonalDetails>
     <Gender>Female</Gender>
     <Name>莫宜端 </Name>
    </PersonalDetails>
   </Reporter>
  </Shot>
 </VideoTrack>
 <AudioTrack UNIT="smpte-val" BEGIN="00:00.00" END="02:01.00">
  <ACoding>
   <AStream>MPEG Audio Layer II</AStream>
   <Sampling UNIT="Hz">44100</Sampling>
   <Channel>Stereo</Channel>
   <BitRate UNIT="kbit/s">192</BitRate>
  </ACoding>
  <SpeakingStyle TYPE="AnchorSession" DIALECT="Cantonese"
  UNIT="smpte-val" BEGIN="00:00.00" END="00:17.00">
   <Anchor UNIT="smpte-val" BEGIN="00:00.00" END="00:06.00">
    <PersonalDetails>
     <Gender>Female</Gender>
     <Name>鄧淑芳 </Name>
    </PersonalDetails>
```

```
</Anchor>
<Anchor UNIT="smpte-val" BEGIN="00:06.00" END="00:17.00">
 <PersonalDetails>
  <Gender>Male</Gender>
  <Name>李燦榮 </Name>
 </PersonalDetails>
</Anchor>
</SpeakingStyle>
<SpeakingStyle TYPE="ReporterSession" DIALECT="Cantonese"
UNIT="smpte-val" BEGIN="00:17.00" END="00:21.00">
 <Reporter>
  <PersonalDetails>
   <Gender>Female</Gender>
   <Name>莫宜端 </Name>
  </PersonalDetails>
 </Reporter>
</SpeakingStyle>
<SpeakingStyle TYPE="NoiseSession" UNIT="smpte-val"
BEGIN="00:21.00" END="00:24.00" />
<SpeakingStyle TYPE="ReporterSession" DIALECT="Cantonese"
UNIT="smpte-val" BEGIN="00:24.00" END="00:47.00">
 <Reporter>
  <PersonalDetails>
   <Gender>Female</Gender>
   <Name>莫宜端 </Name>
  </PersonalDetails>
 </Reporter>
</SpeakingStyle>
<SpeakingStyle TYPE="IntervieweeSession" DIALECT="English"
```

```
UNIT="smpte-val" BEGIN="00:47.00" END="01:08.00">
  <Interviewee>
    <PersonalDetails>
      <Gender>Male</Gender>
    </PersonalDetails>
  </Interviewee>
</SpeakingStyle>
<SpeakingStyle TYPE="ReporterSession" DIALECT="Cantonese"
UNIT="smpte-val" BEGIN="01:08.00" END="01:40.00">
  <Reporter>
    <PersonalDetails>
      <Gender>Female</Gender>
      <Name>莫宜端 </Name>
    </PersonalDetails>
  </Reporter>
</SpeakingStyle>
<SpeakingStyle TYPE="IntervieweeSession" DIALECT="Cantonese"
UNIT="smpte-val" BEGIN="01:40.00" END="01:49.00">
  <Interviewee>
    <PersonalDetails>
    <Gender>Female</Gender>
  </PersonalDetails>
 </Interviewee>
</SpeakingStyle>
<SpeakingStyle TYPE="ReporterSession" DIALECT="Cantonese"
UNIT="smpte-val" BEGIN="01:49.00" END="02:01.00">
  <Reporter>
    <PersonalDetails>
      <Gender>Female</Gender>
```

```
        <Name>莫宜端 </Name>
       </PersonalDetails>
      </Reporter>
    </SpeakingStyle>
  </AudioTrack>
</MEDIA>
<MEDIA TYPE="Text"LocalRef="1999080106.txt" Style="Abstract">
  <TextualInfo>
    <TCoding>
      <Encoding>Unicode</Encoding>
      <Platform>Windows NT/2000/XP</Platform>
    </TCoding>
    <Title UNIT="Byte"BEGIN="1"END="32">
      香港人多近視令一些行業招募有困難
    </Title>
    <Content TYPE="Summary"UNIT="Byte"BEGIN="33"END="184">
      香港年輕一代愈來愈多人有近視，
      而且不少人近視度數都頗深。
      不單令一些行業在招募人員時遇到困難，
      連帶大學內做有關視力的研究時，
      找尋視力完全正常的人也不容易。
    </Content>
    <Content TYPE="Copyright">Television Broadcasts Limited </Content>
  </TextualInfo>
</MEDIA>
</MmML>
```

Figure B.1: An illustration of MmML markup using a news story with filename 1999080409.

Appendix C

Significance Tests

All robust techniques are tested for its experimental significance, the testing procedures are shown below.

C.1 Selection of Cantonese Field Speech Segments

We have performed significance test on Cantonese field speech selection process. We have tested the use of reporter speech only (i.e. retrieval results of anchor speech with reporter speech). We have formulated a paired Z-test to test the significance of the experimental results. The inverse rank obtained for each query with and without fusion of information is $r_w = (r_{w1}, r_{w2}, ..., r_{w1627})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo1627})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w1627} - r_{wo1627})$. Figure C.2 shows the procedures for the significance test on the use of anchor speech with reporter speech using bigrams indexing.

The sample mean of the inverse rank's difference is equals to $\bar{r_d} = 0.008541$ with sample deviation $\sigma_{r_d} = 0.332098$.

The parameter of interest is μ , the mean difference between the inverse rank for each query with and without fusion of information.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.3$

The test statistic is $z_0 = \frac{\bar{r_d} - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject H_0 if $z_0 > z_{0.15} = 1.035$ or if $z_0 < -z_{0.15} = -1.035$.

Since $\bar{r_d} = 0.008541, \sigma_{r_d} = 0.332098$ and $n = 1627$,

$z_0 = \frac{0.008541 - 0}{0.332098/\sqrt{1627}} = 1.037$

Since $z_0 = 1.037 > 1.035$, we reject $H_0 : \mu = 0$ at the 0.3 level of significance.

We conclude that the mean difference between the average inverse rank with and without the use of reporter speech differs from 0. The experiments are performed in a sample of 1,627 experiments.

Figure C.1: A significant test on the use of reporter speech. The experiments are based on bigrams indexing for Cantonese SDR.

C.2 Fusion of Video- and Audio-based Segmentation

We have performed significance test on SDR with fusion of video- and audio-based segmentation. We have formulated a paired Z-test to test the significance of the experimental results. The inverse rank obtained for each query with and without fusion of information is $r_w = (r_{w1}, r_{w2}, ..., r_{w1627})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo1627})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w1627} - r_{wo1627})$. Figure C.2 shows the procedures for the significance test on fusion of video- and audio-based segmentation using bigrams and skipped bigrams indexing based on 1^{st}-best recognition hypothesis only.

C.3 Document Expansion with Reporter Speech

We have performed significance test on SDR with document expansion using reporter speeech. The SDR experiment is performed on the anchor speech from the fusion of video- and audio-based segmentation. We have formulated a paired Z-test to test the significance of the experimental results. The inverse rank obtained for each query with and without fusion of information is $r_w = (r_{w1}, r_{w2}, ..., r_{w1627})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo1627})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w1627} - r_{wo1627})$. Figure C.3 shows the procedures for the significance test on fusion of video- and audio-based segmentation using bigrams and skipped bigrams indexing with document expansion, using extracted reporter speech segments, for Cantonese spoken document retrieval.

The sample mean of the inverse rank's difference is equals to $\bar{r_d} = 0.013046$ with sample deviation $\sigma_{r_d} = 0.190710$.

The parameter of interest is μ , the mean difference between the inverse rank for each query with and without fusion of information.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.01$

The test statistic is $z_0 = \frac{\bar{r_d} - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject H_0 if $z_0 > z_{0.005} = 2.58$ or if $z_0 < -z_{0.005} = -2.58$.

Since $\bar{r_d} = 0.013046, \sigma_{r_d} = 0.190710$ and $n = 1627$,

$z_0 = \frac{0.013046 - 0}{0.190710/\sqrt{1627}} = 2.759$

Since $z_0 = 2.759 > 2.58$, we reject $H_0 : \mu = 0$ at the 0.01 level of significance.

We conclude that the mean difference between the average inverse rank with and without fusion of segmentations differs from 0. The experiments are performed based on 1^{st}-best in a sample of 1,627 experiments.

Figure C.2: A significant test on the fusion of video- and audio-based information. The experiments are based on 1^{st}-best recognition hypothesis using bigrams and skipped bigrams indexing for Cantonese SDR.

The sample mean of the inverse rank's difference is equals to $\bar{r_d} = 0.057468$ with sample deviation $\sigma_{r_d} = 0.28088$.

The parameter of interest is μ , the mean difference between the inverse rank for each query with and without fusion of information.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.01$

The test statistic is $z_0 = \frac{\bar{r_d} - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject H_0 if $z_0 > z_{0.005} = 2.58$ or if $z_0 < -z_{0.005} = -2.58$.

Since $\bar{r_d} = 0.057468, \sigma_{r_d} = 0.28088$ and $n = 1627$,

$z_0 = \frac{0.057468 - 0}{0.28088/\sqrt{1627}} = 8.253$

Since $z_0 = 8.253 > 2.58$, we reject $H_0 : \mu = 0$ at the 0.01 level of significance.

We conclude that the mean difference between the average inverse rank with and without document expansion using reporter speech differs from 0. The test is based on a sample of 1,627 experiments.

Figure C.3: A significant test on the fusion of video- and audio-based information with document expansion. The Cantonese SDR experiment is performed using bigrams and skipped bigrams indexing from the extracted reporter speech segments.

C.4 Document Expansion with N-best Recognition Hypotheses

We have performed significance test on SDR with document expansion using N-best recognition hypotheses. The SDR experiment is performed on the anchor speech from the fusion of video- and audio-based segmentation. We have formulated a paired Z-test to test the significance of the experimental results. The inverse rank obtained for each query with and without fusion of information is $r_w = (r_{w1}, r_{w2}, ..., r_{w1627})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo1627})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w1627} - r_{wo1627})$. Figure C.4 shows the procedures for the significance test on fusion of video- and audio-based segmentation using bigrams and skipped bigrams indexing with document expansion, using N-best recognition hypotheses, for Cantonese spoken document retrieval.

C.5 Document Expansion with Reporter Speech and N-best Recognition Hypotheses

We have performed significance test on SDR with document expansion using reporter speech and N-best recognition hypotheses. The SDR experiment is performed on the anchor speech from the fusion of video- and audio-based segmentation. We have formulated a paired Z-test to test the significance of the experimental results. The inverse rank obtained for each query with and without fusion of information is $r_w = (r_{w1}, r_{w2}, ..., r_{w1627})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo1627})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w1627} - r_{wo1627})$. Figure C.5 shows the procedures for the significance test on fusion of video- and audio-based segmentation using bigrams and skipped bigrams indexing with document expansion, using extracted reporter speech segments and N-best recognition hypotheses,

The sample mean of the inverse rank's difference is equals to $\bar{r_d} = 0.003927$ with sample deviation $\sigma_{r_d} = 0.079544$.

The parameter of interest is μ , the mean difference between the inverse rank for each query with and without fusion of information.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.05$

The test statistic is $z_0 = \frac{\bar{r_d} - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject H_0 if $z_0 > z_{0.025} = 1.96$ or if $z_0 < -z_{0.025} = -1.96$.

Since $\bar{r_d} = 0.003927, \sigma_{r_d} = 0.079544$ and $n = 1627$,

$z_0 = \frac{0.003927 - 0}{0.079544/\sqrt{1627}} = 1.991$

Since $z_0 = 1.991 > 1.96$, we reject $H_0 : \mu = 0$ at the 0.05 level of significance.

We conclude that the mean difference between the average inverse rank with and without document expansion using N-best recognition hypotheses differs from 0. The test is based on a sample of 1,627 experiments.

Figure C.4: A significant test on the fusion of video- and audio-based information with document expansion. The Cantonese SDR experiment is performed using bigrams and skipped bigrams indexing based on N-best recognition hypotheses.

for Cantonese spoken document retrieval.

C.6 Query Expansion with Pseudo Relevance Feedback

We have performed significance test on CLSDR with query expansion using pseudo relevance feedback. We have formulated a paired t-test to test the significance of the experimental results. The average precision obtained from each query batch with and without pseudo relevance feedback is $r_w = (r_{w1}, r_{w2}, ..., r_{w12})$ and $r_{wo} = (r_{wo1}, r_{wo2}, ..., r_{wo12})$ respectively. The difference between the two results sets is $r_d = (r_{w1} - r_{wo1}, r_{w2} - r_{wo2}, ..., r_{w12} - r_{wo12})$. Figure C.6 shows the procedures for the significance test on CLSDR task with query expansion using PRF. The collection is indexed with overlapping character bigrams and skipped bigrams.

The sample mean of the inverse rank's difference is equals to $\bar{r_d} = 0.044401$ with sample deviation $\sigma_{r_d} = 0.131440$.

The parameter of interest is μ , the mean difference between the inverse rank for each query with and without fusion of information.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.01$

The test statistic is $z_0 = \frac{\bar{r_d} - \mu_0}{\sigma_{r_d}/\sqrt{n}}$

Reject H_0 if $z_0 > z_{0.005} = 2.58$ or if $z_0 < -z_{0.005} = -2.58$.

Since $\bar{r_d} = 0.044401, \sigma_{r_d} = 0.19791$ and $n = 1627$,

$z_0 = \frac{0.044401 - 0}{0.19791/\sqrt{1627}} = 9.049$

Since $z_0 = 9.049 > 2.58$, we reject $H_0 : \mu = 0$ at the 0.01 level of significance.

We conclude that the mean difference between the average inverse rank with and without document expansion differs from 0. The expansion is performed using extracted reporter speech segments and N-best recognition

The test it based on a sample of 1,627 experiments.

Figure C.5: A significant test on the fusion of video- and audio-based information with document expansion. Expansion is performed using bigrams and skipped bigrams indexing based on reporter speech segments and N-best recognition hypotheses.

The sample mean of the average precision's difference is equals to $\bar{r_d} = 0.103179$ with sample deviation $s_{r_d} = 0.028315$.

The parameter of interest is μ , the mean difference between the average precision for each query with and without query expansion using PRF.

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

$\alpha = 0.01$

The test statistic is $t_0 = \frac{\bar{r_d} - \mu_0}{s_{r_d}/\sqrt{n}}$

Reject H_0 if $t_0 > t_{0.005,11} = 4.437$ or if $t_0 < -t_{0.005,11} = -4.437$.

Since $\bar{r_d} = 0.103179, s_{r_d} = 0.028315$ and $n = 12$,

$t_0 = \frac{0.103179 - 0}{0.028315/\sqrt{12}} = 12.623$

Since $t_0 = 12.623 > 4.437$, we reject $H_0 : \mu = 0$ at the 0.01 level of significance.

We conclude that the mean difference between the average inverse rank with and without query expansion using PRF differs from 0. The test is based on a sample of 12 query batches.

Figure C.6: A significant test on the experiment with query expansion (by PRF) using overlapping character bigrams and skipped bigrams indexing for CLSDR.

Appendix D

Topic Descriptions of TDT-2 Corpus

The number of stories in each of the 17 topics covered in the CLSDR experiments are listed in Table D.1. The topic ID (TID), title and description of the topics are listed in Table D.2.

Topic ID	Number of Stories	Topic ID	Number of Stories
1	12	48	12
2	12	70	12
5	12	71	12
13	12	76	12
15	12	88	12
20	12	89	3
23	12	91	12
39	12	96	12
44	12		

Table D.1: List of the number of stories in each topic.

TID	Title	Description
1	Asian Economic Crisis	The economic crisis in Asia
2	Monica Lewinsky Case	Allegations about a sexual relationship between Monica Lewinsky and President Clinton
5	Upcoming Philippine Elections	National elections in the Philippines
13	1998 Winter Olympics	1998 Winter Olympic games
15	Current Conflict with Iraq	Saddam Hussein demands a freeze on weapons inspections by the US and UN, while the US and UN demand free access for inspections. Current means the conflict specific to Winter 1998 and its fallout.
20	China Airlines Crash	China Airlines Flight 676 from Bali to Taipei crashes
23	Violence in Algeria	A new wave of bombings and terrorism
39	India Parliamentary Elections	India's Parliamentary Elections
44	National Tobacco Settlement	Devising a National Tobacco Company Settlement
48	Jonesboro shooting	Two students, aged 11 and 13, kill 5 and wound a dozen at their middle school in Jonesboro, Arkansas
70	India – A Nuclear Power?	India begins nuclear testing
71	Israeli-Palestinian Talks (London)	U.S. mediated talks between Israeli and Palestinian leaders occur in London

continue ...

ID	Title	Description
76	Anti-Suharto Violence	Student protests against Indonesian president Suharto based on political differences, motivated by the crushing economic crisis
88	Anti-Chinese Violence in Indonesia	Human rights groups document anti-Chinese violence in Jakarta during riots
89	Afghan Earthquake	Earthquake
91	German Train derails	High speed train derails in Germany
96	Clinton-Jiang Debate	President Clinton and Chinese President Jiang Zemin discuss issues in live televised conference

Table D.2: Topic list of the 17 topics covered in the CLSDR tasks.

Appendix E

Speech Recognition Output from Dragon in CLSDR Task

All Mandarin spoken documents from VOA have been transcribed. Figure E.1 shows an example of the speech recognition output. The story is transcribed using a Chinese CLVSR by Dragon and the output is segmented word units in Simplified Chinese characters (GB coding).

<DOCSET type=ASRTEXT fileid=19980501_0700_0710_VOA_MAN
collect_date=19980501_0700 collect_src=VOA src_lang=MANDARIN
content_lang=NATIVE proc_remarks="Dragon Mandarin ASR">
<W recid=1 Bsec=0.07 Dur=0.47 Clust=4 Conf=0.62>进军
<W recid=2 Bsec=0.54 Dur=0.57 Clust=4 Conf=0.85>新闻
<W recid=3 Bsec=1.11 Dur=0.44 Clust=4 Conf=0.94>提要
<X Bsec=1.55 Dur=1.06 Conf=NA>
<W recid=4 Bsec=2.61 Dur=0.43 Clust=4 Conf=0.94>美国
<W recid=5 Bsec=3.04 Dur=0.64 Clust=4 Conf=0.98>国务卿
<W recid=6 Bsec=3.68 Dur=0.84 Clust=4 Conf=0.99>奥尔布莱特
<W recid=7 Bsec=4.53 Dur=0.45 Clust=4 Conf=0.77>呼吁
<W recid=8 Bsec=5.01 Dur=0.45 Clust=4 Conf=0.96>有关
<W recid=9 Bsec=5.46 Dur=0.15 Clust=4 Conf=0.93>各
<W recid=10 Bsec=5.61 Dur=0.33 Clust=4 Conf=0.91>方
<W recid=11 Bsec=5.94 Dur=0.32 Clust=4 Conf=0.93>继续
<W recid=12 Bsec=6.26 Dur=0.32 Clust=4 Conf=0.80>在
<W recid=13 Bsec=6.58 Dur=0.26 Clust=4 Conf=0.89>新
<W recid=14 Bsec=6.84 Dur=0.44 Clust=4 Conf=0.85>展开
<W recid=15 Bsec=7.28 Dur=0.35 Clust=4 Conf=0.91>外交
<X Bsec=7.63 Dur=0.76 Conf=NA>
<W recid=16 Bsec=8.39 Dur=0.34 Clust=4 Conf=0.39>越南
<W recid=17 Bsec=8.73 Dur=0.37 Clust=4 Conf=0.79>北韩
<W recid=18 Bsec=9.10 Dur=0.11 Clust=4 Conf=0.93>的
<W recid=19 Bsec=9.21 Dur=0.43 Clust=4 Conf=0.89>同意
<X Bsec=9.64 Dur=1.11 Conf=NA>
<W recid=20 Bsec=10.75 Dur=0.36 Clust=1 Conf=0.53>希望
<W recid=21 Bsec=11.11 Dur=0.19 Clust=1 Conf=0.86>在

continue ...

<W recid=22 Bsec=11.30 Dur=0.32 Clust=1 Conf=0.21>来到

<W recid=23 Bsec=11.62 Dur=0.45 Clust=1 Conf=0.92>西藏

<W recid=24 Bsec=12.07 Dur=0.45 Clust=1 Conf=0.91>精神

<W recid=25 Bsec=12.52 Dur=0.48 Clust=1 Conf=0.96>领袖

<W recid=26 Bsec=13.00 Dur=0.60 Clust=1 Conf=0.95>达赖喇嘛

<W recid=27 Bsec=13.60 Dur=0.45 Clust=1 Conf=0.95>表示

<X Bsec=14.05 Dur=0.69 Conf=NA>

<W recid=28 Bsec=14.74 Dur=0.35 Clust=1 Conf=0.84>从

<W recid=29 Bsec=15.09 Dur=0.30 Clust=1 Conf=0.93>经济

<W recid=30 Bsec=15.39 Dur=0.35 Clust=1 Conf=0.85>角度

<W recid=31 Bsec=15.74 Dur=0.34 Clust=1 Conf=0.92>讲

<X Bsec=16.08 Dur=0.31 Conf=NA>

<W recid=32 Bsec=16.39 Dur=0.58 Clust=1 Conf=0.93>西藏

<W recid=33 Bsec=16.97 Dur=0.47 Clust=1 Conf=0.77>继续

<W recid=34 Bsec=17.44 Dur=0.35 Clust=1 Conf=0.96>作为

<W recid=35 Bsec=17.79 Dur=0.35 Clust=1 Conf=0.97>中国

<W recid=36 Bsec=18.14 Dur=0.12 Clust=1 Conf=0.92>的

<W recid=37 Bsec=18.26 Dur=0.61 Clust=1 Conf=0.70>一部分

<X Bsec=18.87 Dur=0.34 Conf=NA>

<W recid=38 Bsec=19.21 Dur=0.28 Clust=1 Conf=0.81>会

<W recid=39 Bsec=19.49 Dur=0.31 Clust=1 Conf=0.80>得到

<W recid=40 Bsec=19.80 Dur=0.06 Clust=1 Conf=0.63>的

<W recid=41 Bsec=19.86 Dur=0.59 Clust=1 Conf=0.96>好处

<W recid=42 Bsec=20.53 Dur=0.37 Clust=1 Conf=0.70>也许

<W recid=43 Bsec=20.90 Dur=0.20 Clust=1 Conf=0.88>更

<W recid=44 Bsec=21.10 Dur=0.33 Clust=1 Conf=0.88>多

continue ...

<X Bsec=21.43 Dur=1.41 Conf=NA>
<W recid=45 Bsec=22.84 Dur=0.40 Clust=4 Conf=0.96>美国
<W recid=46 Bsec=23.24 Dur=0.36 Clust=4 Conf=0.97>国会
<W recid=47 Bsec=23.60 Dur=0.63 Clust=4 Conf=0.97>参议院
<W recid=48 Bsec=24.23 Dur=0.46 Clust=4 Conf=0.93>投票
<W recid=49 Bsec=24.69 Dur=0.40 Clust=4 Conf=0.94>批准
<W recid=50 Bsec=25.09 Dur=0.20 Clust=4 Conf=0.96>了
<W recid=51 Bsec=25.29 Dur=0.46 Clust=4 Conf=0.45>接纳
<W recid=52 Bsec=25.75 Dur=0.56 Clust=4 Conf=0.77>波兰
<W recid=53 Bsec=26.39 Dur=0.63 Clust=4 Conf=0.60>匈牙利
<W recid=54 Bsec=27.05 Dur=0.29 Clust=4 Conf=0.92>和
<W recid=55 Bsec=27.34 Dur=1.24 Clust=4 Conf=0.97>捷克共和国
<W recid=56 Bsec=28.64 Dur=0.37 Clust=4 Conf=0.83>为
<W recid=57 Bsec=29.01 Dur=0.52 Clust=4 Conf=0.89>北约
<W recid=58 Bsec=29.53 Dur=0.63 Clust=4 Conf=0.97>成员国
<W recid=59 Bsec=30.16 Dur=0.10 Clust=4 Conf=0.92>的
<W recid=60 Bsec=30.26 Dur=0.40 Clust=4 Conf=0.94>计划
<X Bsec=30.66 Dur=1.63 Conf=NA>
<W recid=61 Bsec=32.29 Dur=0.42 Clust=6 Conf=0.95>下面
<X Bsec=32.71 Dur=0.12 Conf=NA>
<W recid=62 Bsec=32.83 Dur=0.30 Clust=6 Conf=0.91>请
<W recid=63 Bsec=33.13 Dur=0.32 Clust=6 Conf=0.87>听
...
</DOCSET>

Figure E.1: An example of speech recognition output of a Mandarin news story with filename VOA19980501.0700.0036.

Appendix F

Parameters Estimation

We have performed a few sets of CLSDR experiments based on overlapping character bigrams indexing. There are four sets of experiments in total, they are used for the estimation of the parameters: N_r, N_{rt}, N_n and N_{nt}.

F.1 Estimating the Number of Relevant Documents, N_r

We have fixed the values of the parameters, N_{rt}, N_n and N_{nt}, to ten, zero and zero respectively. N_{rt} is picked up randomly while N_n and N_{nt} are set to zero so that the effect of N_r on retrieval performance can be seen clearly. A set of CLSDR experiments have been carried out using different values of N_r and the results are shown in Table F.1.

N_r	0	1	2	3	4	5
mAP	0.410	0.475	**0.477**	0.469	0.457	0.443

Table F.1: The retrieval performance (in mAP) based on different values of N_r.

Based on the retrieval performance in Table F.1, we set N_r to two in all the CLSDR experiments.

F.2 Estimating the Number of Terms Added from Relevant Documents, N_{rt}, to Original Query

We have fixed the values of the parameters, N_r, N_n and N_{nt}, to two, zero and zero respectively. N_r is set to two based on the result from the previous Section while N_n and N_{nt} are set to zero so that the effect of N_{rt} on retrieval performance can be observed clearly. A set of CLSDR experiments have been carried out using different values of N_{rt} and the results are shown in Table F.2.

N_{rt}	0	10	20	30	40	50	60	70
mAP	0.410	0.477	0.491	0.499	0.502	0.505	0.507	0.505
N_{rt}	80	90	100	110	120	130	140	
mAP	0.507	0.510	0.512	0.512	**0.513**	0.512	0.512	

Table F.2: The retrieval performance (in mAP) based on varies values of N_{rt}.

Based on the retrieval performance in Table F.2, we set N_{rt} to 120 for in all the CLSDR experiments.

F.3 Estimating the Number of Non-relevant Documents, N_n, from the Bottom-scoring Retrieval List

We have fixed the values of the parameters, N_r, N_{rt} and N_{nt}, to zero, zero and ten respectively. N_{nt} is chosen randomly while N_r and N_{rt} are set to zero so that the effect of N_{rt} on retrieval performance can be noticed quickly. A set of CLSDR experiments have been carried out using different values of N_n and the results are shown in Table F.3.

Based on the retrieval performance in Table F.3, we set N_n to one for in all the CLSDR experiments.

N_n	0	1	2	3	4	5
mAP	0.410	**0.411**	0.410	0.410	0.410	0.410

Table F.3: The retrieval performance (in mAP) based on diverse values of N_n.

F.4 Estimating the Number of Terms, Selected from Non-relevant Documents (N_{nt}), to be Removed from Original Query

We have fixed the values of the parameters, N_r, N_{rt} and N_n, to two, a hundred and twenty and one respectively. The numbers are based on the results from previous experiments. A set of CLSDR experiments have been carried out using different values of N_{nt} and the results are shown in Table F.4.

N_{rt}	0	10	20	30	40	50
mAP	0.51255	0.51256	0.51249	0.51328	0.51348	**0.51363**
N_{rt}	60	70	80	90	100	
mAP	0.51361	0.51355	0.51357	0.51359	0.51360	

Table F.4: The retrieval performance (in mAP) based on varied values of N_{rt}.

Based on the retrieval performance in Table F.4, we set N_{nt} to twenty for in all the CLSDR experiments.

Appendix G

Abbreviations

Table G.1 includes abbreviations that occur in this thesis for quick reference.

AIR	Average Inverse Rank
AoE-IT	Area of Excellence in Information Technology
AP	Average Precision
APW	Associated Press Worldstream Service
ASR	Automatic Speech Recognition
BLA	Between Language Ambiguity
CLSDR	Cross-language Spoken Document Retrieval
CMU	Carnegie Mellon University
CN	Cosine Normalization
DLN	Document Length Normalization
FCM	Fuzzy C-means
fps	Frame per Second
FVAS	Fusion of video- and audio-based segmentation
GMM	Gaussian Mixture Models
GTC	Graph-theoretical Cluster
HDM	Histogram Difference Metric
HMM	Hidden Markov Model
IDF	Inverse Document Frequency
IR	Information Retrieval
KIR	Known-item Retrieval
LSHK	Linguistic Society of Hong Kong
LVCSR	Large-vocabulary Continuous Speech Recognition

mAP	Mean Average Precision
MEI	Mandarin-Engish Information
MFCC	Mel-frequency Cepstral Coefficients
MLSDR	Multilingual Spoken Document Retrieval
MmML	Multimedia Markup Language
MPEG	Moving Picture Experts Group
MST	Minimum Spanning Tree
NSC	Non-significant Change
NYT	New York Times Newswire Service
OOV	Out-of-vocabulary
PRF	Pseudo Relevance Feedback
QE	Query Expansion
SC	Significant Change
SDM	Spatial Difference Metric
SDR	Spoken Document Retrieval
SMIL	Synchronized Multimedia Integration Language
TDT	Topic Detection and Tracking
TF	Term Frequency
TID	Topic ID
TREC	Text REtrieval Conference
TVB	Television Broadcasts Limited
VOA	Voice of America
VSM	Vector Space Model
WLA	Within-language Ambiguity
XML	eXtensible Markup Language

Table G.1: A list of abbreviations used in this thesis.

Bibliography

[1] OYEZ^{TM} U.S. Supreme Court Multimedia. http://www.oyez.org/oyez/frontpage.

[2] History and Politics Out Loud (HPOL): a searchable archive of politically significant audio materials. http://www.hpol.org.

[3] Survivors of the Shoah Visual History Foundation. http://www.vhf.org.

[4] OASIS: Organization for the Advancement of Structured Information Standards http://www.oasis-open.org/home/index.php.

[5] Palowitch, C. and D. Stewart. Automating the Structural Markup Process in the Conversion of Print Documents to Electronic Text. *Digital Libraries 1995: The Second Annual Conference on the Theory and Practice of Digital Libraries*, 1995. [online: http://www.csdl.tamu.edu/DL95/papers/palowitc/palowitc.html]

[6] Oard, D. W. and B. J. Dorr. A Survey of Multilingual Text Retrieval Technical Report. In *Technical Report UMIACS-TR-96-19 CS-TR-3615 of the Electronic Engineering Department*, University of Maryland, USA, 1996.

[7] Morrison, P. and P. Morrison. Wonders: The Sum of Human Knowledge? Article from *Scientific American*, July, 1998.

[8] Hauptmann, A. G., P. Scheytt, H. D. Wactlar and P. E. Kennedy. Multilingual Informedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages. In *Proceedings of the DARPA Workshop on Broadcast News Understanding Systems*, Lansdowne, USA, February, 1998.

[9] Alis Technologies and the Internet Society. Web Languages Hit Parade. http://babel.alis.com/palmares.en.html, June, 1997.

[10] Source: Global Reach. http://global-reach.biz/globstats/evol.html.

[11] NIST Special Publication 500-240: The Sixth Text REtrieval Conference (TREC-6). Gaithersburg, Maryland, November, 1997. http://trec.nist.gov/pubs/trec6/t6_proceedings.html.

[12] Text REtrieval Conference (TREC). http://trec.nist.gov.

[13] Topic Detection and Tracking (TDT). http://www.nist.gov/TDT.

[14] SpeechBot^{TM}– audio search using speech recognition. http://speechbot.research.compaq.com/.

[15] Ordelman, R., A. V. Hessen, F. D. Jong. Lexicon Optimization for Dutch Speech Recognition in Spoken Document Retrieval. In *Proceedings of* 7^{th} *European Conference on Speech Communication and Technology*, pages 1085–1088, Aalborg, Denmark, September, 2001.

[16] Larson, M., S. Eickeler, G. Paass, E. Leopold and J. Kindermann. Exploring Sub-Word Features and Linear Support Vector Machines for German Spoken Document Classification. In *Proceedings of* 7^{th} *International Conference on Spoken Language Processing*, pages 1989–1992, Denver, USA, September, 2002.

[17] Shah, C. and A. N. Khan. Spoken Document Retrieval (SDR) for Broadcast Bews in Indian Languages In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, May, 2001.

[18] Nishizaki, H. and S. Nakagawa. Comparing Isolately Spoken Keywords with Spontaneously Spoken Queries for Japanese Spoken Document Retrieval. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1505–1508, Denver, USA, September, 2002.

[19] Furui, S., K. Ohtsuki, Z. P. Zhang. Japanese Broadcast News Transcription and Information Extraction. *Communications of the ACM*, pages 71–73, Volume 43, Issue 2, February, 2000.

[20] Meinedo, H. and J. Neto. Automatic Speech Annotation and Transcription in a Braadcast News Task. In *Proceedings of the ISCA Workshop on Multilingual Spoken Document Retrieval*, pages 95–100, Hong Kong, China, April, 2003.

[21] Chien, L. F., H. M. Wang, B. R. Bai, S. C. Lin. A Spoken Access Approach for Chinese Text and Speech Information Retrieval. *Journal of American Society for Information Science (JASIS)*, pages 313–323, Volume 51, Issue 4, 2000.

[22] Wang, H. M. and B. Chen. Content-based Language Models for Spoken Document Retrieval. *International Journal of Computer Processing of Oriental Languages*, pages 193–209, Volume 14, Issue 2, June, 2001.

[23] Li, Y. C., H. Meng, W. K. Lo and P. C. Ching. Multi-scale audio indexing for Chinese spoken document retrieval. In *Proceedings of the International Conference on Spoken Language Processing*, pages 101–104, Beijing, China, October 2000.

[24] Informedia™Project at School of Computer Science, Carnegie Mellon University. http://www.informedia.cs.cmu.edu/.

[25] Wactlar, H., A. Olligschlaeger, A. G. Hauptmann and M. Christel. Complementary Video and Audio Analysis for Broadcast News Archives. *Communications of the ACM*, pages, 42–47, Volume 43, Issue 2, February, 2000.

[26] Hauptmann, A. G., T. D. Ng and R. Jin. Video Retrieval Using Speech and Image Information. In *Proceedings of the Electronic Imaging Conference, Storage and Retrieval for Multimedia Databases*, Santa Clara, USA, January, 2003.

[27] Hauptmann, A. G., D. Lee, P. E. Kennedy. Topic Labeling of Multilingual Broadcast News in the Informedia Digital Video Library. *Joint ACM Digital Library/SIGIR Workshop on Multilingual Information Discovery and AccesS (MIDAS)*, Berkeley, USA, August, 1999.

[28] Rough'n'Ready™ audio indexing system. http://www.bbn.com/speech/roughnready.html.

[29] Colbath, S. and F. Kubala. Rough'n'Ready: A Meeting Recorder and Browser. In *Proceedings of the Workshop on Perceptual User Interfaces*, San Francisco, USA, November, 1998.

[30] Meng, H., B. Chen, E. Grams, S. Khudanpur, G. Levow, W. K. Lo, D. Oard, P. Schone, H. M. Wang and J. Wang. Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval. In *Technical Report for Johns Hopkins University Summer Workshop 2000*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA, 2000.

[31] Sheridan, P., M. Wechsler and P. Schauble. Cross Language Speech Retrieval: Establishing a Baseline Performance. In *Proceedings of the 20th*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 99–108, Philadelphia, USA, July, 1997.

[32] Fujii, A., K. Itou and T. Ishikawa. LODEM: A Multilingual Lecture-on-Demand System. In *Proceedings of ISCA Workshop on Multilingual Spoken Document Retrieval*, pages 13–18, Hong Kong SAR, China, April, 2003.

[33] LIMSI-CNRS: Processing multilingual broadcast audio for information access. http://www.limsi.fr/tlp/audioindex.html.

[34] Garofolo, J. S., G. P. Auzanne and E. M. Voorhees. The TREC Spoken Document Retrieval Task: A Success Story. In *Proceedings of the Recherche d'Informations Assistee par Ordinateur: Content-Based Multimedia Information Access Conference*, pages 107–126, Paris, France, April, 2000.

[35] Schauble, P. and M. Wechsler. First Experiences with a System for Content Based Retrieval of Information from Speech Recordings. In *Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval*, pages 59–69, Montreal, Canada, August, 1995.

[36] Witbrock, M. J. and A. G. Hauptmann. Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents. In *Proceedings of the DARPA Speech Recognition Workshop*, Chantilly, USA, February, 1997.

[37] Ng, K. and V. Zue. *Subword-based Approaches for Spoken Document Retrieval.* Ph.D. Thesis, Department of Electronic Engineering and Computer Science, Massachusetts Institute of Technology, USA, February, 2000.

[38] Brown, M., J. Foote, G. Jones and S. Young. Open Vocabulary Speech Indexing for Voice and Video Mail Retrieval. In *Proceedings of the Fourth*

ACM International Conference on Multimedia , pages 307–316, Boston, USA, November, 1996.

[39] Kwok, K. L. Comparing Representations in Chinese Information Retrieval. In *Proceedings of the* 20^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, Philadelphia, USA, July, 1997.

[40] Taniguchi, Y., A. Akutsu, Y. Tonomura and H. Hamada. An Intuitive and Efficient Access Interface to Real-time Incoming Video based on Automatic Indexing. In *Proceedings of the Third ACM International Conference on Multimedia* , pages 25–33, San Francisco, USA, November, 1995.

[41] Merlino, A., D. Morey and M. Maybury. Broadcast News Navigation using Story Segmentation. In *Proceedings of the Fifth ACM International Conference on Multimedia*, pages 381–391, Seattle, USA, November, 1997.

[42] Kobla, V., D. Doermann and D. Faloutsos. Video Trails: Representing and Visualizing Structure in Video Sequences. In *Proceedings of the Fifth ACM International Conference on Multimedia*, pages 335–346, Seattle, USA, November, 1997.

[43] Hauptmann, A. G. and M. J. Witbrock. Story Segmentation and Detection of Commercials in Broadcast News Video. In *Proceedings of Advances in Digital Libraries Conference*, pages 168–179, Santa Barbara, USA, April, 1998.

[44] Virtual Tutorials in Phonology – Hong Kong Word. http://www.cbs.polyu.edu.hk/VTP/hkword/s/s1.htm.

[45] Chien, L. F.. Fast and quasi-natural language search for gigabits of Chinese texts. In *Proceedings of the 18th ACM SIGIR Conference on Research*

and Development in Information Retrieval, pages 112–120, Seattle, USA, July, 1995.

[46] Kwok, K. L.. Lexicon Effects on Chinese Information Retrieval. In *Proceedings of the Second Conference on Empirical Methods in NLP*, ACL, pages 141–148, Providence, USA, August 1997.

[47] Chen, B., H. M. Wang, and L. S. Lee. Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan using Syllable-Level Statistical Characteristics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 3, pages 1771–1774, Istanbul, Turkey, June, 2000.

[48] Li, Y. C., W. K. Lo, H. Meng and P. C. Ching. Query Expansion using Phonetic Confusion for Chinese Spoken Document Retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pages 89–93, Hong Kong SAR, China, October, 2000.

[49] Area of Excellence in Information Technology – Multimedia Repository. http://www.se.cuhk.edu.hk/~aoe/.

[50] TVB News. http://news.tvb.com/tvnews/index.html.

[51] Real.com. http://www.real.com/.

[52] International Organisation for Standardisation – Coding of Moving Pictures and Audio.
http://mpeg.telecomitalialab.com/standards/mpeg-1/mpeg-1.htm.

[53] Synchronized Multimedia Integration Language (SMIL 2.0).
http://www.w3.org/TR/smil20.

[54] SMIL Media Object Modules: Definition of Continuous Media. http://www.w3.org/TR/smil20/extended-media-object.html#media-Definitions.

[55] Lo, W. K., H. M. Meng and P. C. Ching. Sub-syllabic Acoustic Modeling across Chinese Dialects. In *Proceedings of the Second International Symposium on Chinese Spoken Language Processing*, pages 97–100, Beijing, China, October, 2000.

[56] Lo, W. K., T. Lee and P. C. Ching. Development of Cantonese Spoken Language Corpora for Speech Applications. In *Proceedings of International Symposium on Chinese Spoken Language Processing*, pages 102–107, Singapore, December, 1998.

[57] Linguistic Society of Hong Kong. *Hong Kong Jyut Ping Character Table*, Linguistic Society of Hong Kong, Hong Kong SAR, China, 1997.

[58] Lo, W. K., T. Lee and P. C. Ching. CULEX^{TM}and CUPDICT^{TM} *Technical Report for Department of Electronic Engineering*, Digital Signal Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China, 1999.

[59] Li, Y. C.. *The Use of Subword-based Audio Indexing in Chinese Spoken Document Retrieval.* M. Phil. Thesis, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China, 2001.

[60] Salton, G. and M. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hall, New York, New Jersey, USA, 1983.

[61] Yeung, M. M. and B. L. Yeo. Time-constrained Clustering for Segmentation of Video into Story Units. In *Proceedings of the IEEE 13th Interna-*

tional Conference on Pattern Recognition, pages 375–380, Vienna, Austria, August 1996.

[62] Hui, P. Y., X. Tang, H. Meng, W. Lam and X. Gao. Automatic Story Segmentation for Spoken Document Retrieval. In *Proceedings of the* 10^{th} *IEEE International Conference on Fuzzy Systems*, Melbourne, Australia, December, 2001.

[63] Hui, P. Y., W. K. Lo and H. Meng. Multimedia Fusion in Automatic Extraction of Studio Speech Segments for Spoken Document Retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong SAR, China, April, 2003.

[64] Meng, H., X. Tang, P. Y. Hui and X. Gao. Speech retrieval with video parsing for television news programs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1401–1404, Salt Lake City, USA, May 2001.

[65] Gao, X. and X. Tang. Automatic Parsing of News Video based on Cluster Analysis. In *Proceedings of the 2000 Asia Pacific Conference on Multimedia Technology and Applications*, Kaohsiung, Taiwan, December, 2000.

[66] Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithm*, Plenum, New York, 1981.

[67] Zahn, C. T. Graph-theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, pages 68–86, Volume 20, Number 1, 1971.

[68] Chen, T., C. Huang, E. Chang and J. C. Wang. Automatic Accent Identification using Gaussian Mixture Models. In *Proceedings of the Automatic Speech Recognition and Unierstanding*, pages 343–346, Trento, Italy, 2001.

[69] Reynolds, D. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communications*, pages 91–108, Volume 17, Elsevier Science, 1995.

[70] Singhal, A. and F. Pereira. Document Expansion for Speech Retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41, Berkeley, USA, August, 1999.

[71] Hui, P. Y., W. K. Lo and H. Meng. Two Robust Methods for Cantonese Spoken Document Retrieval. In *Proceedings of the ISCA Workshop on Multilingual Spoken Retrieval*, Macau / Hong Kong SAR, China, April, 2003.

[72] Ballesteros, Lisa Anne. *Resolving Ambiguity for Cross-language Information Retrieval: a Dictionary Approach.* Ph. D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, USA, 2001.

[73] Uzuner, O., B. Katz and D. Yuret. Word Sense Disambiguation for Information Retrieval. In *Proceedings of the 1999 16th National Conference on Artificial Intelligence*, Orlando, USA, July, 1999.

[74] Sanderson, M.. Word Sense Disambiguation and Information Retrieval. Ph. D. Thesis, Department of COmputing Science, University of Glasgow, UK, 1997.

[75] Fung, P. and K. McKeown. A Technical Word and Term Translation Aid using Noisy Parallel Corpora across Language Groups. *Machine Translation*, pages 53–87, 1996.

[76] Kraaij, W. and D. Hiemstra. Cross Language Retrieval with the Twenty-one System. In *Proceedings of tge Sixth Retrieval Conference*, pages 753–760, Gaithersburg, 1997.

[77] Dagan, I., A. Itai and U. Schwall. Two Languages are More Informative than One. In *Proceedings of the* 29^{th} *Annual Meeting of the Association for Computational Linguistics*, pages 130–137, 1991.

[78] Linguistic Data Consortium, Project Topic Detection and Tracking Phase Two (TDT-2). http://www.ldc.upenn.edu/Projects/TDT2.

[79] New York Times on the Web. http://www.nytimes.com/.

[80] Voice of America – Chinese. http://www.voanews.com/chinese/.

[81] BBN Technologies – Identifinder^{TM}. http://www.bbn.com/speech/identifinder.html.

[82] ScanSoft – Dragon LVCSR. http://www.dragonsys.com/.

[83] Singhal, A., C. Buckley and M. Mitra. Pivoted Document Length Normalization. In *Research and Development in Information Retrieval*, pages 21–29, 1996.

[84] Zhang, H. J., Z. Chen, W. Y. Liu and M. J. Li. Relevance Feedback in Content-Based Image Search. Invited Keynote. In *Proceedings of* 12^{th} *International Conference on New Information Technology (NIT)*, Beijing, China, May, 2001.

[85] Hoashi, K., E. Zeitler and N. Inoue. Implementation of Relevance Feedback for Content-based Music Retrieval Based on User Preferences. In *Proceedings of the* 25^{th} *Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR*, pages 385–386, Tampere, Finland, August, 2002.

[86] Hoashi, K., K. Matsumoto, N. Inoue and K. Hashimoto. Document Filtering Method using Non-relevant Information Profile. In *Proceedings of*

the 25th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 176–183, Athens, Greece, July, 2000.

[87] Kwok, K. L., N. Dinstl and P. Deng. English-Chinese CLIR using a Simplified PIRCS System. In *Proceedings of the First International Conference on Human Language Technology Research (HLT)*, pages 87–90, San Diego, USA, March, 2001.

[88] Rocchio, J. J.. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 313–323, Prentice Hall, 1971.

[89] Porkaew K., K. Chakrabarti and S. Mehrotra. Query Refinement for Multimedia Similarity Retrieval in MARS. In *Proceedings of the 8th Annual ACM International Conference on Multimedia*, Orlando, USA, November, 1999.