Multi-scale spoken document retrieval for Cantonese broadcast news

Correspondence:	Wai Kit Lo
Affiliation:	Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.
Email:	wklo @ ieee.org
Tel.:	852-3163 4073
Fax:	852-2603 5505
Address:	Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

Multi-scale spoken document retrieval for Cantonese broadcast news

Wai-Kit Lo, Helen M. Meng, and P. C. Ching

(wklo@ieee.org, hmmeng@se.cuhk.edu.hk, pcching@ee.cuhk.edu.hk)

The Chinese University of Hong Kong, Shatin,
Hong Kong SAR, China

Abstract: This paper presents the application of a multi-scale paradigm to Chinese spoken document retrieval (SDR) for improving retrieval performance. Multi-scale refers to the use of both words and subwords for retrieval. Words are basic units in a language that carry lexical meaning and subword units (such as phonemes, syllables or characters) are building components for words. Retrieval using subword indexing units is found to perform better than words because of the robustness of subword units to out-of-vocabulary (OOV) words during speech recognition and ambiguities in word segmentation. Experimental results have demonstrated that subword bigrams can bring improvement in retrieval performance over words (~9.56%). Application of multi-scale fusion to SDR aims at combining the lexical information of words and the robustness of subwords. This work presents the first detailed investigation for a Cantonese broadcast news retrieval task using two different multi-scale fusion approaches: pre-retrieval fusion and post-retrieval fusion. Multi-scale retrieval using both words and syllable bigrams achieve improvement in retrieval performance (~1.90%) over retrieval on the composite scales.

Index Terms: speech retrieval, multi-scale fusion, Cantonese speech recognition

1 Introduction

The popularity of the World Wide Web (WWW) has attracted a large amount of information available over the Internet. In order to retrieve information relevant to user queries from the WWW effectively, search technologies are being actively developed and improved. With the advent of multimedia technology, there is also a large amount of information available in *multimedia* formats (e.g. recordings from TV or radio broadcast, presentations, meetings or lectures). This has brought about the demand for content-based retrieval from multimedia information sources. Among these multimedia sources, speech is a common information channel. Therefore, effective location of information from speech recordings is important for searching multimedia data. Spoken document retrieval (SDR) is the technology that enables efficient search of information that is relevant to user queries from collection of speech recordings. Applications of SDR include content-based multimedia document organization, surveillance for security or intelligence purposes, education and entertainment software, etc.

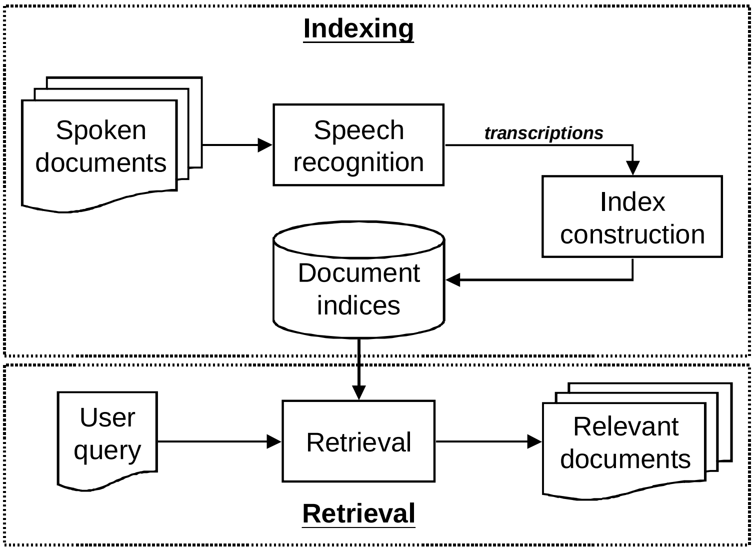


Figure 1 Overview of the document indexing and retrieval processes in a spoken document retrieval system.

A common approach for SDR is to combine automatic speech recognition (ASR) and textual information retrieval (IR) as shown in Figure 1. ASR is applied to produce textual transcriptions for spoken documents. Indices for these spoken documents are then built using the automatic transcriptions. Based on the document indices, textual IR is applied to retrieve documents that are relevant to the user's query. Traditionally, words are used for building document indices in information retrieval. Use of the word units is beneficial to retrieval because words carry lexical meaning. However, a major problem of word-based retrieval for SDR is the existence of words in the spoken documents that are unknown to the speech recognizer (out-of-vocabulary or OOV). OOV words are either missed or replaced by other in-vocabulary words. Furthermore, ASR may introduce recognition errors. Document indexing with these combined errors will lower the retrieval performance. Since all words are composed from subwords units (e.g. phonemes, syllables, or characters), retrieval based on subword indexing units can be applied for improving retrieval performance when there are OOV words.

This paper reports on the first investigation of SDR for Cantonese broadcast news. Cantonese is a Chinese dialect used in Hong Kong, Macau and south China areas. Since Cantonese has different vocabulary and grammar in its spoken and written form, this imposes additional difficulties to retrieval from ASR transcriptions based on textual queries. Spoken documents used in this work also make up the first collection of Cantonese broadcast news in Internet multimedia format (RealMedia^{TM}) for SDR experiments. We propose to apply a multi-scale paradigm to SDR for improving performance of retrieval from imperfect document indices. Multi-scale refers to the use of both words and subwords for retrieval (Meng et. al. 2000a). We have also investigated two approaches for multi-scale fusion: *pre-retrieval* fusion and *post-retrieval* fusion. Extensive experiments have also been conducted for investigations that also demonstrate consistent retrieval performance improvements obtained.

This paper is organized as follows. In Section 2, we will have a review of some background information and previous work in related areas. Section 4 presents the details of two approaches for multi-scale retrieval: pre-retrieval fusion and post-retrieval fusion. Our experimental setup is given in Section 4. Section 5 gives result and analyses for our experiments. Finally, this work is concluded in Section 6.

2 Background

2.1 Spoken document retrieval

Spoken document retrieval (SDR) has attracted much recent interest (Schauble, 1997). As shown in Figure 1, essential processes involved are document *indexing* and *retrieval* based on user queries. Indexing in SDR involves converting speech data into transcriptions by automatic speech recognition. Document indices are then derived from these transcriptions by the indexing procedure. When a user supplies a query to the retrieval system, the retrieval process will search through the document indices for relevant documents. Relevant documents are returned to the user as retrieval results.

There are several pioneering systems in SDR. The Video Mail Retrieval (Foote, 1997) achieves audio and video document retrieval by the keyword-spotting algorithm. Another example is the Infomedia digital video library project (Hauptmann, 1997) that aims to offer effective search and retrieval of relevant information from digital library. SDR has evolved from the early keyword-spotting based retrieval (Foote, 1997 and Jones et. al., 1996b) to more recent systems based on large scale broadcast news transcription systems working in conjunction with information retrieval systems (Thong et. al., 2000, Makhoul et. al., 2000, and Johnson et. al., 2001). The focus has been placed on improving the accuracy of automatic speech recognition to reduce errors in transcriptions and hence document indices. With the availability of large-scale broadcast news archives, many retrieval systems have been developed for automatic indexing and retrieval of broadcast news. To name a few examples, these include the Broadcast News Navigator from Mitre (Merlino and Maybury, 1999), SCAN from AT&T (Choi el. al, 1998 and Whittaker et. al., 1999), SpeechBot of Compaq (Thong et. al., 2000 and Compaq, 2000), Rough’n Ready of BBN (Makhoul et. al., 2000), and Multimedia Document Retrieval project of Cambridge University (Johnson et. al., 2001 and Tuerk et. al., 2001). However, highly accurate automatic speech recognition is not always feasible. Under many situations, the acoustic condition is harsh and the automatic recognition accuracy is unavoidably low. Alternative approaches have to be used for improving retrieval performance under these conditions. This has motivated the investigation of multi-scale SDR in this work.

2.2 Spoken document retrieval for Chinese

Chinese language is significantly different from Indo-European languages (such as English, German, etc.). For example, Chinese is character-based and English is alphabetical. Among the dialects of Chinese, Mandarin (or Putonghua) is the official language and Cantonese is a dialect that is used in Hong Kong, Macau and south China areas. Between Mandarin and Cantonese, they share the same basic phonological structure in that both of them are monosyllabic and tonal. Monosyllabic means that every character is pronounced as a syllable. Every Chinese syllable consists of an optional syllable initial (a consonant) together with a syllable final (a vowel followed by an optional consonant). Chinese syllables are tonal because syllables with different lexical tones have different meanings. In Chinese languages, lexical tones are expressed acoustically in pitch profiles and temporal durations of syllable finals. Table 1 shows a comparison of the phonological statistics for Cantonese and Mandarin. It should be noted that from the different initials, finals, and tones in each dialect, the number of valid syllables among all possible combinations is limited (~1800 out of 6360 possible combinations for Cantonese and ~1200 out of 4025 for Mandarin).

Table 1 Comparison of phonological statistics between Cantonese and Mandarin.

	Cantonese	**Mandarin**
No. of syllables (without tone)	~600	~400
No. of tonal syllables	~1800	~1200
No. of lexical tones	6	5
No. of initials	19	22
No. of finals	53	35

In the written form, words in Chinese text appear as a sequence of characters. All Chinese textual materials are formed form about 10,000 different traditional Chinese characters.[i] Moreover, there is no explicit word delimiter in written Chinese. As a result, word segmentation process is needed for extraction of words from written Chinese. As shown in Figure 2, potential word units are identified from a character sequence based on a given vocabulary of words. Sequences of consecutive characters that can form words are extracted. For those characters that cannot be segmented as word, they are returned as separated characters.

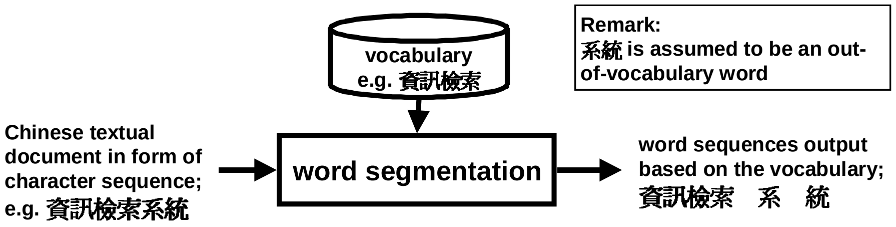
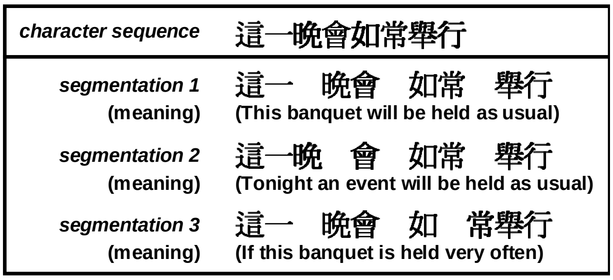

vocabulary
e.g. 資訊檢索
Remark:
系統 is assumed to be an out-of-vocabulary word
Chinese textual document in form of character sequence;
e.g. 資訊檢索系統
word segmentation
word sequences output based on the vocabulary;
資訊檢索 系 統

Figure 2 Illustration of the word extraction process from text and the potential problem. Since word segmentation is dependent on pre-defined vocabulary, it is susceptible to the problem of out-of-vocabulary (OOV) words. In this example, the character sequence for the Chinese word “information retrieval system” is to be segmented. If the word “system” is an OOV word, the composite character will be returned as separated characters.

Moreover, word segmentation from character sequence is an ambiguous process. For example, there can be three different segmentations for the character sequence “這一晚會如常舉行” as shown in Figure 3. These segmentations are all syntactically valid and semantically meaningful. If an incorrect segmentation is used in the document indexing process, irrelevant documents will be returned from the retrieval.



character sequence	這一晚會如常舉行
segmentation 1 (meaning)	這一 晚會 如常 舉行 (This banquet will be held as usual)
segmentation 2 (meaning)	這一晚 會 如常 舉行 (Tonight an event will be held as usual)
segmentation 3 (meaning)	這一 晚會 如 常舉行 (If this banquet is held very often)

Figure 3 Illustration of ambiguity in word segmentation. A character sequence can be segmented into different sequences of words that are syntactically valid and semantically meaningful.

Extraction of word units from spoken documents can be achieved by application of speech recognizers such as large vocabulary continuous speech recognizer (LVCSR) or keyword-spotters for spoken documents. Words obtainable in these cases are then dependent on the vocabulary of the speech recognizers. Any word in the speech signal not covered by the vocabulary (i.e. out-of-vocabulary word) may either be treated as noise or substituted by other words contained in the vocabulary. This is illustrated with an example in Figure 4.

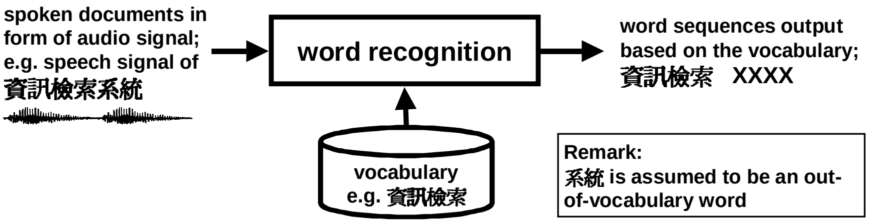

spoken documents in form of audio signal; e.g. speech signal of 資訊檢索系統
word recognition
word sequences output based on the vocabulary; 資訊檢索 XXXX
vocabulary e.g. 資訊檢索
Remark: 系統 is assumed to be an out-of-vocabulary word

Figure 4 Illustration of the word extraction process from speech and the potential problem. Speech recognition is applied to obtain word units from speech signal. If the word "system" is an OOV word, it is either substituted by other words or missed by the recognizer. This is denoted as "XXXX" in the output.

Early work in Chinese SDR includes those by (Chen et. al., 2000a, Chen et. al., 2000b, Wang, 2000, Chen et. al., 2001a, and Chen et. al., 2001b, Chen et. al. 2002) that work on the retrieval of Mandarin broadcast news. Investigations for retrieval of Cantonese broadcast news include those in (Li et. al, 2000, Meng et. al., 2000, Meng et. al., 1999, and Meng and Hui, 2001). With the availability of large amounts of data for Chinese,[ii] there have also been some investigations on cross-language SDR for Chinese such as the retrieval of Mandarin audio documents using English queries in the MEI project (Meng et. al., 2000b, Lo et. al., 2001, Meng et. al., 2001a, Meng et. al., 2001b, and Wang et. al, 2001).

2.3 *Subword-based retrieval*

Words are made up from subwords where subwords can be phonemes, syllables or characters in Chinese. Subword units for indexing can be obtained by forming *overlapping n-grams* from subword sequence. An important characteristic of subword n-grams is that the formation process is independent of any vocabulary. One may also select different values of *n* to change the degree of the *sequential contextual information* captured by subword n-grams. Sequential contextual information is the information captured by the consecutive sequence of subword units in the n-gram. In order to reduce the possibility of missing any information contained in the subword sequence, *overlapping* subword n-grams are usually used. As an example, assume that there is a sequence of *m* subword units (can be phonemes, syllables, or characters, etc.)

$$\{S_1\ S_2\ S_3 \ldots S_m\},$$

the sequences of overlapping n-grams for *n* = 2 and *n* = 3 are

$$\{\ S_1S_2\ \ S_2S_3\ \ S_3S_4\ \ldots\ S_{m-1}S_m\ \} \qquad \text{for } n = 2,$$

$$\{\ S_1S_2S_3\ \ S_2S_3S_4\ \ldots\ S_{m-2}S_{m-1}S_m\ \} \qquad \text{for } n = 3.$$

Retrieval using subword-based indexing units has the advantage that the indexing process is not affected by the OOV problem. As shown in Figure 2 and Figure 4, the word extraction process relies on the word coverage of the vocabulary. In case there are OOV words (e.g. new proper names or acronyms etc.), they cannot be extracted and hence cannot contribute to retrieval. Retrieval using subword n-grams is more robust to the OOV problem since the formation of subword n-grams is independent of any pre-defined vocabulary.

In addition, subword n-grams are also robust to ambiguities in word segmentation. As shown in Figure 3, there are multiple ways of segmenting a Chinese character sequence. In case the segmentation process returns an incorrect word sequence for indexing, incorrect documents will be retrieved. By using overlapping subword n-grams (e.g. bigrams), a unique sequence of indexing units can be obtained (see Figure 5). This can avoid the problem due to ambiguities in word segmentation.

character sequence	這一晚會如常舉行
character bigram	這一 一晚 晚會 會如 如常 常舉 舉行

Figure 5 Formation of a unique set of character bigrams from a given character sequence for indexing.

Subword n-grams are also robust to some errors introduced by speech recognition as illustrated below. On the word scale, a single mis-recognized syllable/character would render the word match to fail. If subword n-grams are used instead, some correct n-grams can be preserved for retrieval. For example, assume that the original 4-character word is C_1 C_2 C_3 $\mathbf{C_4}$. If the speech recognition process introduced a single error making C_4 become E_4, the character sequence becomes C_1 C_2 C_3 $\mathbf{E_4}$. As a result, the word C_1 C_2 C_3 C_4 cannot contribute to the retrieval process. If character bigrams are used, three character bigrams can be formed from this word. The recognition error in this example only cause one of these targeted character bigrams (the last in the word) to be substituted by an erroneous character bigram as shown below. The two correct bigrams C_1C_2 and C_2C_3 remains correct and can contribute to retrieval.

C_1C_2 C_2C_3 $C_3\mathbf{C_4}$	original bigram sequence
C_1C_2 C_2C_3 $C_3\mathbf{E_4}$	bigram sequence with the error

In addition, the use of subword syllable indexing units is also beneficial when there are recognition errors due to homophones. Recognition errors due to homophones are possible when words are substituted by incorrect words having same syllable pronunciations. Under such circumstance, some of the output characters are wrong but the syllable sequence is correct (see Figure 6). As a result, retrieving with character-based units will miss these words. Assume that there are two homophones $\{C_1C_2\}$ and $\{C_{1a}C_2\}$ and their pronunciations are both $\{S_1S_2\}$. If $\{C_1C_2\}$ is an OOV word in the recognition vocabulary, any occurrence of this word in the speech signal will be substituted by the homophone $\{C_{1a}C_2\}$. Therefore, the correct word cannot be matched during retrieval using character-based units. If syllable-based units are used, the indexing unit $\{S_1S_2\}$ can be preserved for retrieval.

	Character	Syllable	Meaning
Out-of-vocabulary word	估價	**/gu-gaa/**	**valuation**
Word in vocabulary	股價	**/gu-gaa/**	**stock price**

Figure 6 Homophones introduce erroneous character-based indexing units (1 of the 2 characters is incorrect). In this example, the out-of-vocabulary word is substituted by a homophone in the vocabulary. The incorrect unit cannot contribute to the retrieval process.

Recognition errors due to homophones are especially common for foreign proper names because there are new names introduced continuously. More importantly, their Chinese names are often phonetic transliterations - translation based on the pronunciation. As a result, speech recognizers may return different character sequences with the same pronunciations. As shown in Figure 7, the OOV word, Vajpayee, is substituted by a character sequence with the same syllable pronunciation. Neither the expected word nor character bigrams can be obtained from the recognition output. As a result, the recognition error cause retrieval based on character-based units to fail.

	Character	Syllable
Out-of-vocabulary word	雅杰帕伊	/ngaa-git-paak-ji/
Recognizer output	雅 傑 拍 衣	/ngaa/ /git/ /paak/ /ji/

Figure 7 Proper names not covered by the vocabulary (e.g. Vajpayee) are represented as separated characters with same pronunciations as the targeted ones. Retrieval on either character bigram or word scale will miss this name. Use of syllable-based indexing units can avoid such error.

There has been some work for English using subword phoneme n-grams as indexing units (Ng, 1998, Ng and Zue, 1998, and Ng, 2000b). In (Wechsler, 2000), there is also investigation on the use of subword phonemes for retrieving spoken documents in German and English. For information retrieval in Chinese, there has been some work on word segmentation for Chinese IR (Nie and Brisebois, 1996) as well as investigation on the use of words or character n-grams as indexing units for Chinese textual IR (Kwok, 1997 and Nie and Ren, 1999). Subword-based indexing is useful for information retrieval in Chinese because all textual documents are made up of a finite number of characters. Using character n-grams in Chinese IR guarantees full textual coverage of documents. In addition, problems due to ambiguity in word segmentation can be avoided. Similarly, the use of syllable n-grams offers full phonological coverage for Chinese languages. Similar to English mentioned above, the use of subword indexing units in Chinese SDR can also overcome the OOV problem in speech recognition. Improvement in Chinese SDR performance by using subword indexing units has been reported for Cantonese in (Meng et. al., 1999, Li et. al., 2000, Meng et. al., 2000a) as well as Mandarin in (Wang et. al., 2000). Previous experiences in subword-based information retrieval have demonstrated that the optimal value of *n* for n-grams depends on the language. The use of phoneme 4-grams (Ng, 2000b) in English and subword bigrams (*n*=2, character bigrams and syllable bigrams) for Chinese (Kwok, 1997, Meng et. al., 2000a, and Wang 2000) achieves the best retrieval performance among other values of *n*.

2.4 Multi-scale approach

In this work, we investigate the fusion of multiple indexing scales (words and subwords) in SDR for improving retrieval performance. Multi-scale refers to the use of both words and subwords. Words are lexically oriented units that carry discrete piece of information. Matching pairs of words usually refer to the same idea. However, words are made up of subword units (such as phonemes, characters or syllables). Matching subword units may come from different words. Therefore, words have greater *specificity* and are more discriminative for the retrieval process than subwords. On the other hand, subword scale indexing units are usually sequential n-grams of subword units that may not have lexical meaning. The advantage of using subword indexing units is that full coverage is guaranteed and hence they are more robust to the OOV problem than words.

Multi-scale retrieval aims at fusing the *specificity* of words and the *robustness* of subwords to obtain improvement in retrieval performance. Early work in information retrieval that makes use of multiple information sources include fusing retrieval results from multiple retrieval systems as reported in (Fox et. al., 1992, Fox and Shaw, 1993, and Bartell et. al., 1994), using multiple index sources for document indexing (Jones et. al., 1996), and combining retrieval results of different phoneme n-grams by linear combination (Ng 2000a, Ng, 2000b, and Ng 2000c). Retrieval performance after fusion has been demonstrated to be improved over individual retrieval results.

Multi-scale units for Chinese include words and subword n-grams in characters or syllables. Characters are lexically-based and syllables are phonetically-oriented units. Figure 8 shows examples of the indexing units formed on word and subword scale in character and syllable for the Chinese word “information retrieval”.

	characters	*syllables*
words	資訊檢索 (word)	/zi_seon_gim_sok/ (word in syllables)
subwords	資訊 , 訊檢 & 檢索 (e.g. character bigrams)	/zi_seon/, /seon_gim/ & /gim_sok/ (e.g. syllable bigrams)

1) 資訊檢索 means information retrieval
2) Syllables are shown as Cantonese transcription

Figure 8 Different indexing units for Chinese on word and subword scales in forms of characters and syllables.

For Chinese languages, the use of fused retrieval results has also be been reported in (Nie and Ren 1999) for textual retrieval as well as Cantonese SDR in (Meng et. al., 1999, and Meng et. al., 2000a) and Mandarin SDR in (Chen et. al., 2000a, and Chen et. al. 2001a). Among these previous experiences, both words and subword n-grams (characters or syllables) have been used in multi-scale retrieval and obtained improvements in retrieval performance. In this work, we attempt to present a thorough investigation and analysis of multi-scale fusion for Chinese SDR and specifically for the retrieval of Cantonese broadcast news under lower recognition accuracy.

3 Multi-scale fusion

Fusion of multiple indexing scales is proposed for improving retrieval performance by taking advantage of different indexing scales. In general, multi-scale fusion can be carried out in two different ways: *pre-retrieval fusion* and *post-retrieval fusion.*

In pre-retrieval fusion, the query and document representations on different scales are merged *before retrieval*. This enables all of the information in the representations to be used during retrieval. Pre-retrieval fusion enlarges the dimensional space of indexing units and this allows more information to be represented. However, in case there is a change in the fusion process (e.g. addition of more scales or change of weightings for composite scales), the whole document collection will have to be indexed again. These can induce significant overhead to the overall retrieval process.

In post-retrieval fusion, fusion is carried out *after retrieval*. Post-retrieval fusion allows us to merge the retrieval results from different scales as well as different retrieval engines. The advantage of this approach is its simplicity. For example, change of weightings for composite scales does not require the document collection to be indexed again. Furthermore, the dimensional space of indexing units do not increase. In case only the retrieval scores are accessible from a retrieval system, post-retrieval fusion is still possible but not for pre-retrieval fusion. The major drawback of post-retrieval approach is that additional retrieval runs have to be performed.

Given the representation of query and document on scale *k* be q^k and d^k respectively, the retrieval process returns retrieval scores on scale *k* as

$$Score\left(q^k, d^k\right) = D\left(q^k, d^k\right) \qquad (1)$$

where *D(q,d)* is the retrieval process returning a score between the query and document vectors q^k and d^k on scale *k*.

3.1 Pre-retrieval multi-scale fusion

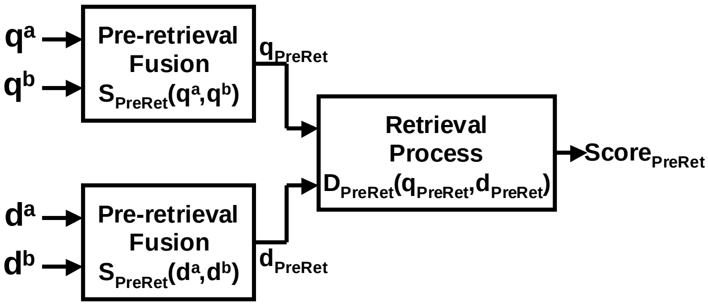


Figure 9 Block diagram of the pre-retrieval multi-scale fusion process.

Pre-retrieval multi-scale fusion merges document representations on composite scales before retrieval. The same process is also applied to query representations as illustrated in Figure 9. Without loss of generality, we assume that representations on scale *a* and scale *b* are *m* and *n* dimension respectively. The merging function is then a relation that maps the *m* and *n* dimension vectors to a *p* dimension representation,

$$S_{\text{PreInt}}: \Re^{m} \text{ x } \Re^{n} \rightarrow \Re^{p},$$

where *m* and *n* are dimensions of individual scales before fusion and *p* is the dimension of the representation after fusion. $p \leq m+n$ if the merging is a simple weighted concatenation. In (2), the overall fusion and retrieval process is shown.

$$Score_{\text{PreInt}}\left(q^{a}, q^{b}, d^{a}, d^{b}\right) = D_{\text{PreInt}}\left(S_{\text{PreInt}}\left(q^{a}, q^{b}\right), S_{\text{PreInt}}\left(d^{a}, d^{b}\right)\right) \quad (2)$$

In our experiments, weighted concatenation of the individual scales is used as the pre-retrieval fusion function. Other fusion methods are also possible. For example, words from LVCSR are augmented by new words from a word spotter in (Jones et. al., 1996). The fusion process used for our pre-retrieval fusion experiments is shown in (5). Weighting factors for scale *a* and *b* are w_a and w_b respectively and it is constrained that $w_a + w_b = 1$. Given the query and document vectors for scale *a* as follows.

$$q^{a} = \left[q_{1}^{a} \; q_{2}^{a} \cdots q_{m}^{a}\right] \quad (3)$$

$$d^{a} = \left[d_{1}^{a} \; d_{2}^{a} \cdots d_{m}^{a}\right] \quad (4)$$

The fused representation for the query is

$$\begin{aligned} S_{\text{PreInt}}\left(q^{a}, q^{b}\right) &= \left[w_{a} \cdot q^{a} \quad w_{b} \cdot q^{b}\right] \\ &= \left[w_{a} \cdot q_{1}^{a} \; w_{a} \cdot q_{2}^{a} \cdots w_{a} \cdot q_{m}^{a} \quad w_{b} \cdot q_{1}^{b} \; w_{b} \cdot q_{2}^{b} \cdots w_{b} \cdot q_{m}^{b}\right] \end{aligned} \quad (5)$$

The same process is also applied to the document vector

$$\begin{aligned} S_{\text{PreInt}}\left(d^{a}, d^{b}\right) &= \left[w_{a} \cdot d^{a} \quad w_{b} \cdot d^{b}\right] \\ &= \left[w_{a} \cdot d_{1}^{a} \; w_{a} \cdot d_{2}^{a} \cdots w_{a} \cdot d_{m}^{a} \quad w_{b} \cdot d_{1}^{b} \; w_{b} \cdot d_{2}^{b} \cdots w_{b} \cdot d_{m}^{b}\right] \end{aligned} \quad (6)$$

By applying (5) and (6) to (2), we can calculate retrieval scores as shown in (7). The ranked list of documents for the given queries can then be obtained based on these scores.

$$Score_{\mathrm{PreInt}}\left(q^{a}, q^{b}, d^{a}, d^{b}\right) = D_{\mathrm{PreInt}}\Big(\big\lfloor w_a \cdot q_1^a \; w_a \cdot q_2^a \cdots w_a \cdot q_m^a \quad w_b \cdot q_1^b \; w_b \cdot q_2^b \cdots w_b \cdot q_m^b \big\rfloor, \\ \big[w_a \cdot d_1^a \; w_a \cdot d_2^a \cdots w_a \cdot d_m^a \quad w_b \cdot d_1^b \; w_b \cdot d_2^b \cdots w_b \cdot d_m^b \big]\Big) \tag{7}$$

3.2 Post-retrieval multi-scale fusion

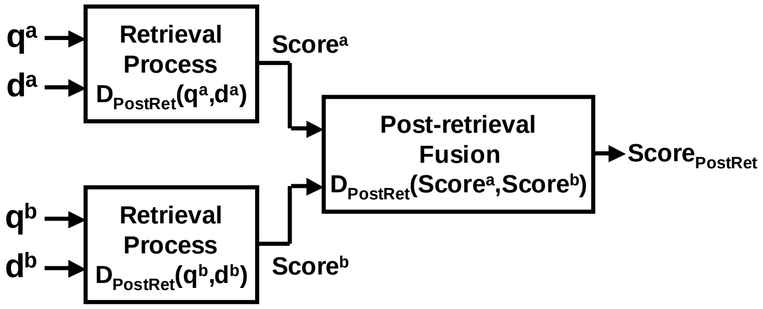


Figure 10 Block diagram of the post-retrieval multi-scale fusion process.

Post-retrieval multi-scale fusion is achieved by merging the retrieval results from different indexing scales to obtain new rankings for documents in the collection. Figure 10 shows the post-retrieval fusion process that merges retrieval scores from different indexing scales to obtain new retrieval scores. Documents are then re-ranked according to the new retrieval scores.

Post-retrieval fusion is a mapping function for two real numbers to a new one,

$$S_{\mathrm{PostInt}}: \Re^1 \text{ x } \Re^1 \to \Re^1.$$

Given retrieval scores from two scales *a* and *b*, the fusion process is to merge these values into a new value

$$Score_{\mathrm{PostInt}}\left(q^{a}, q^{b}, d^{a}, d^{b}\right) = S_{\mathrm{PostInt}}\left(Score^{a}, Score^{b}\right) \tag{8}$$

By applying (1) to (8) for post-retrieval fusion, we have

$$Score_{\mathrm{PostInt}}\left(q^{a}, q^{b}, d^{a}, d^{b}\right) = S_{\mathrm{PostInt}}\left(D_{\mathrm{PostInt}}\left(q^{a}, d^{a}\right), D_{\mathrm{PostInt}}\left(q^{b}, d^{b}\right)\right) \tag{9}$$

where D_{PostInt} is the retrieval process that returns a retrieval score.

Post-retrieval multi-scale fusion provides another perspective to the fusion process. Results of the composite runs are merged. In this work, a weighted linear combination is adopted and the fusion function is given below

$$Score_{\mathrm{PostInt}} = \sum_{k=a,b} w^{k} \cdot D_{\mathrm{PostInt}}\left(q^{k}, d^{k}\right) \tag{10}$$

where w_k is the weight for scale *k* and it is constrained that $\sum w^k = 1$ for simplicity.

4 Experimental setup

This section describes details of the experimental setup for multi-scale retrieval. These include details of the collection of audio-video clips used as our spoken documents, configuration of the large vocabulary continuous speech recognizer (LVCSR) used for automatic transcription, information retrieval model and weighting schemes used in our experiments, as well as our task formulation and performance evaluation criterion.

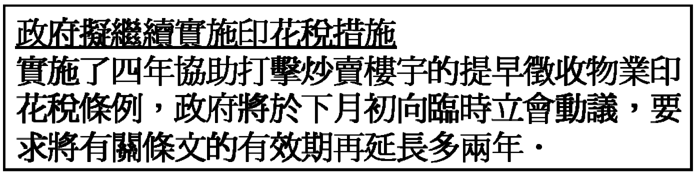
4.1 *Audio-video data for retrieval*

The collection of spoken documents used in the retrieval experiments consists of 1800 audio-video clips of television news collected over the Internet. This makes up the first collection of Cantonese broadcast news data for spoken document retrieval purpose. News clips in this collection are evening news program from a television company in Hong Kong. Every news clip contains a single news story. It has a short description from the anchor that is usually followed by detailed report from reporters and interviews. This collection of news data spans the period from June 1997 to February 1998. The total duration of video is 54 hours. Table 2 shows the summary of statistics for these data.

Table 2 Summary of statistics for the 1800 news stories in the data archive used for our retrieval experiments.

	Average	**Range**
Audio track (seconds)	84.12	79 – 1419
Textual summary (characters)	70.83	10 – 270
News headline (characters)	12.89	4 – 39

From these news clips, the audio tracks[iii] are extracted for automatic transcription and the video tracks are left untouched. Since these news clips are made available over the Internet, the audio tracks in the clips are heavily compressed using a CELP based codec to 8.5kbps. This has imposed additional difficulty to the automatic transcription process.


政府擬繼續實施印花稅措施

實施了四年協助打擊炒賣樓宇的提早徵收物業印花稅條例，政府將於下月初向臨時立會動議，要求將有關條文的有效期再延長多兩年·

Figure 11 An example of the news summary that provides a gist of the news story. The headline is underlined on the top as shown.

Every news clip has a short textual summary and a headline. Figure 11 shows an example textual summary and the headline is underlined. The headline is a highlight of the news story. The textual summary is essentially the script for the anchor that gives more details of the news story. In most cases, the textual summary does not repeat the headline. It should be noted that textual summaries cannot be treated as the transcriptions of the audio tracks due to several reasons. First, the summary is in written form and the anchors always rephrase the content when speaking. Moreover, the spoken and written form of Cantonese has different grammar and vocabulary. The most important difference between audio tracks and textual summaries is that there are recordings from reporters and interviewees in addition to descriptions from anchors.

In this work, the extracted audio tracks are used as our spoken document collection and textual summaries are used as the textual document collection. The headlines are used as query for our retrieval experiments.

4.2 *Large vocabulary continuous speech recognizer*

A large vocabulary continuous speech recognizer (LVCSR) is developed for obtaining automatic transcriptions from audio tracks of news clips. This recognizer operates using 16-mixture left-and-right context-dependent initial-final models (Lo et. al., 2000), together with a 41k word bigram language model.

The acoustic models are trained using 25 hours of transcribed news from the same Internet source and the training data is excluded from the document collection for retrieval. The language model is trained using a 274Mbyte (192,230 articles) newspaper archive covering the period from March 1997 to February 1998. All articles are segmented into words using a left-to-right maximum matching algorithm together with the 41k lexicon (CULEX). A word bigram language model is then estimated from the segmented texts with Good-Turing discounting using CMU-Cambridge SLM Toolkit (Clarkson et. al., 1997). Outputs of the LVCSR are sequences of words.

Evaluation of the LVCSR has been made using a non-overlapping evaluation set containing 0.5 hour of transcribed news data. Table 3 shows the recognition performance of this LVCSR.

Table 3 The base syllable and character recognition accuracy for the LVCSR. (Note: Cantonese is a tonal language, base syllable refers to the syllable identifies without the tone information.)

	LVCSR	
	Base syllable	Character
accuracy	45.43%	38.07%

4.3 *Information retrieval*

4.3.1 Multi-scale query and document processing

Queries and documents are processed into different indexing scales for our retrieval experiments. Textual headlines and news summaries received the same processing while additional speech recognition process is applied to spoken documents to obtain textual representations.

Figure 12 shows the extraction of indexing units on different scales for both the textual materials (headlines and summaries) and LVCSR outputs. For textual materials, character sequences are segmented into word sequences and overlapping n-grams are formed. On syllable scales, words in syllables are obtained by pronunciation lookup (based on CULEX) for the sequences of words. Syllable n-grams are then derived from the sequence of words in syllables. For the spoken documents, they are first automatically transcribed using the Cantonese LVCSR. The word sequences output from the LVCSR are then processed into different scales. Transcriptions of spoken documents received the same processing as the textual materials except that word segmentation is not needed.

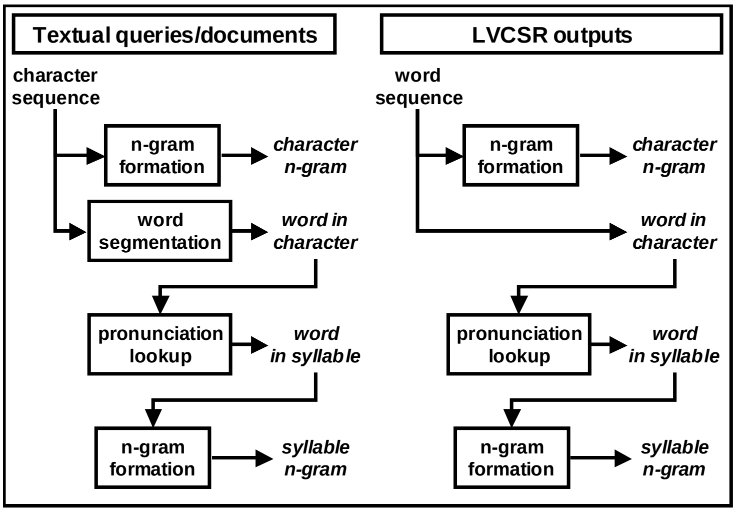


Figure 12 Formation of indexing units on different scales for the two sources of documents.

4.3.2 Retrieval model

Retrieval experiments in this paper are performed using the vector space model (VSM). In a VSM, queries and documents are represented as vectors.

$$q = [q_1 \; q_2 \cdots q_n] \tag{11}$$

$$d = [d_1 \; d_2 \cdots d_n] \tag{12}$$

where q and d are vectors with elements q_i and d_i respectively. Value of the ith element is the weight of the ith indexing units.

In this work, weighting functions for query and document are defined as

$$q_i = (\log(tf_{q_i}) + 1.0) \cdot \log\left(\frac{N+1}{n_i}\right) \tag{13}$$

$$d_i = \log(tf_{d_i}) + 1.0 \tag{14}$$

where tf_{qi} is the frequency of the ith indexing unit in the query vector and tf_{di} is the frequency of the ith indexing unit in the document vector. N is the total number of documents in the collection and n_i is the number of documents in the collection containing the ith indexing unit.

In (13) and (14), logarithm of frequencies are used to avoid too much emphasis on frequent units. The $\log\left(\frac{N+1}{n_i}\right)$ term in (13) is an implementation of *inverse document frequency* (IDF). IDF is applied to down-weigh units that appear across many documents because these units are non-discriminative and not useful for retrieval.

Retrieval scores obtained from the VSM is based on a similarity measure between the query and document vectors. The cosine similarity measure defined as shown in (15) is used in this work.

$$Score(q,d) = \frac{q \cdot d}{\|q\| \cdot \|d\|} \tag{15}$$

4.3.3 Task formulation and evaluation measure

The 1800 news clips and textual summaries are used as our document collections. Since our document collection does not have topic relevance judgment, we have formulated a known-item-retrieval (KIR) task for our retrieval experiments. The aim of a KIR task is to retrieve a specific document for each query. In our retrieval experiments, each of the headlines is used as query to retrieve the corresponding document.

Performance evaluation for KIR task is achieved using the average inverse rank (AIR). AIR is effectively the same as mean average precision where there is only *one* relevant document. In a KIR task, it is the specific document to be retrieved. As shown in (16), AIR is obtained by taking average over the inverse of the ranks of the targeted documents for every query. The range of AIR lies between 0 and 1. An AIR of 1 means that all targeted documents are ranked at the top in the ranked retrieval lists.

$$AIR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{16}$$

where N is the total number of query-document pairs, $rank_i$ is the rank of the ith document when retrieved using the ith query.

5 Results and analysis

5.1 *Retrieval on individual scales*

In order to investigate the retrieval performance of indexing units on different scales, KIR experiments have been carried out using textual summaries and LVCSR outputs as document collections. Retrievals using textual summaries can provide textual IR performance reference for comparison to SDR performance obtained using LVCSR outputs. These retrieval experiments are carried out using words and subword n-grams in forms of characters and syllables.

5.1.1 Results based on textual retrieval

The processed news headlines and textual summaries are used for retrieval experiments on word and subword scales in forms of characters and syllables. Overlapping n-grams are used as the subword scale indexing units and the value of *n* ranges from one (unigram) to five.

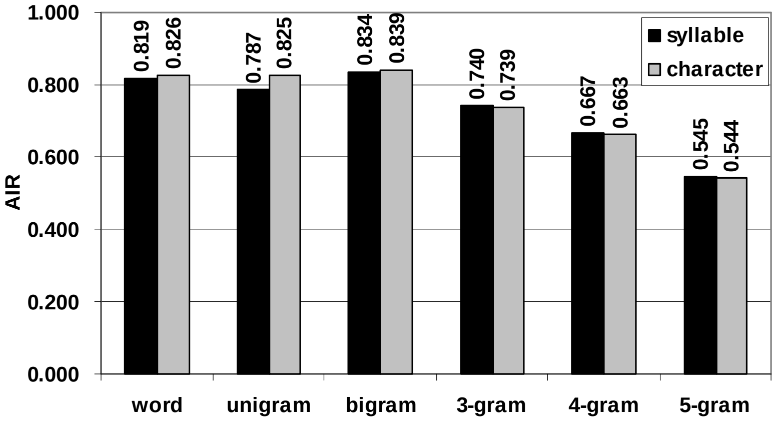


Figure 13 Retrieval results for documents derived from textual summaries and indexed on different scales.

It is found that subword bigrams perform the best among other indexing scales for retrieval from textual summaries using headlines. As shown in Figure 13, retrieval performances in AIR are 0.834 and 0.839 when using syllable bigrams and character bigrams respectively. It can also be seen that as the value of *n* increases from one to five for the subword n-grams, retrieval performance peaks at bigram scale and gradually drop from bigram to 5-gram. This can be explained by the fact that there are not many words having four or five characters. The majority of words are two characters long. This can be seen from the statistics of the Cantonese pronunciation lexicon (CULEX) in Table 4. Among all of the polysyllabic words, over 75% of them are two characters long. As a result, using long n-grams as indexing units reduces the chance of getting matches because words shorter than *n* characters may be missed during retrieval. Therefore, retrieval using subword bigrams performs better than longer n-grams.

Table 4 Percentage count of polysyllabic words at different lengths in the pronunciation lexicon.

Word length	2	3	4	5	…
Word count	76%	13%	10%	0.89%	…

In addition, it is also observed that retrievals using subword n-grams as indexing units perform better than retrieval using words. This performance gain is obtained because the coverage of vocabulary is incomplete. Out-of-vocabulary (OOV) words cause matches on the word scale to fail. Since subword bigrams are independent of any vocabulary, matching units can still be found for OOV words. Furthermore, according to the statistics in Table 4, the majority of the Chinese words are two characters long. As a result, retrieval using subword bigrams is robust to the OOV problem. OOV words may be new words (e.g. people names, company names, etc.) that are particularly common in the news domain. In Figure 14, an example extracted from the experimental output that illustrates the advantage of using subword bigrams when there is OOV word (underlined).

Query:
word :八 佰 伴 員工 向 工 聯會...
bigram :八佰 佰伴 伴員 員工 工向 向工 工聯 ...

Document:
word :八 佰 伴 整個 清盤 過程 要...
bigram :八佰 佰伴 伴整 整個 個清 清盤 盤過 ...

Figure 14 This example shows that out-of-vocabulary word encountered during retrieval will cause match on word scale to fail. Retrieval using character bigrams is not affected by the OOV problem.

If we compare the retrieval performances in Figure 13 for different indexing scales in forms of syllables and characters, there is no big difference except for unigrams. When the indexing units change from syllable- to character-based, character unigrams perform better than syllable unigrams because there is a large amount of homophones. As shown in Table 5, there are around 10,000 characters corresponding to around 600 syllables. This effectively reduces the maximum number of distinct indexing units. When syllable-based indexing units are used, the effect of homophones renders these units less discriminative and hence lowers the retrieval performances. When the length of subword n-grams increases, the differences in performances between syllable- and character-based indexing units decrease. It is because discrimination power of long n-grams is dominated by the sequential contextual information. In addition, a long unit is less likely to have homophone than a unigram. For example, it is shown in Table 5 that there are altogether 318 5-character words that correspond to 315 different syllable pronunciations. The number of homophones decreases significantly as the word length increases. As a result, this makes the degradation more prominent on unigram scales than longer subword n-grams.

Table 5 Statistics of base syllable homophones for words in the pronunciation lexicon.

Word length	Word count	No. of different pronunciations	Ratio
1	around 10,000	around 600	16.67
2	27,057	22,108	1.22
3	4,720	4,682	1.01
4	3,579	3,525	1.02
5	318	315	1.01

5.1.2 Results of the Cantonese SDR task

When performing our Cantonese SDR experiments, word sequences output from the LVCSR are used for deriving indexing units on different scales – words and subword n-grams in forms of syllables and characters. SDR experiments are then carried out over these scales and results are shown in Figure 15.

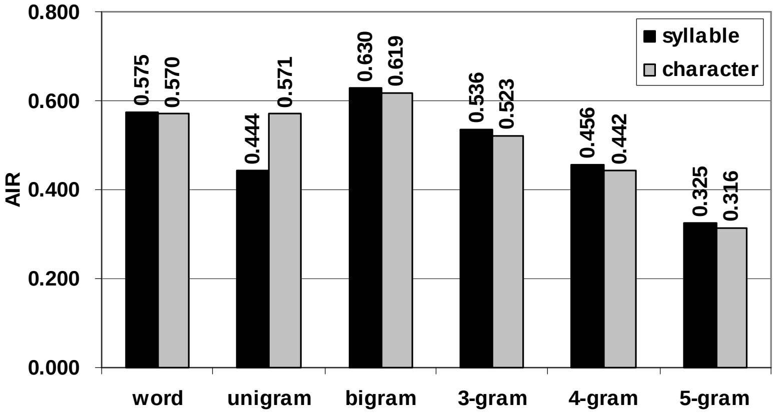


Figure 15 Cantonese SDR results for documents derived from LVCSR outputs and indexed on different scales.

Our results show that subword bigrams perform the best among other indexing scales in Cantonese SDR. This is consistent with the observation from retrieving textual summaries. Robustness of subword bigrams also contributes to the performance gain over words. Since there are recognition errors in automatic transcriptions of spoken documents, the chance of having long sequences of correctly recognized characters is smaller than two characters units. Therefore, use of subword bigrams for Cantonese SDR can cover some of the mis-recognized words. Furthermore, retrieval using subword bigrams can also avoid problems due to discrepancies between the word identified from segmentation of the textual queries and the words returned by LVCSR.

Results also show that there are only small differences in retrieval performances between syllable- and character-based indexing units. This is also consistent with retrieval from textual summaries. Since indexing units on all scales used in our SDR experiments are derived from the same LVCSR output, differences in retrieval performance are introduced by the difference in the length of the n-grams. As mentioned before, the discrimination power of subword n-grams relies more on the sequential contextual information. As a result, there are only minor differences in performances between syllable- and character-based subword n-grams except for unigrams. This is because unigrams contain no contextual information and the effect of homophone dominates.

Our SDR experimental results also show that syllable-based units perform marginally better than character-based units. Syllable bigrams and words in syllables have AIR of 0.630 and 0.575 respectively while character bigrams and words achieve AIR of 0.619 and 0.570 respectively. We found that these performance differences are introduced by recognition errors due to homophones. If there are character recognition errors in transcriptions due to homophone, those erroneous words or character n-grams cannot contribute to retrieval. By using syllable-based indexing units, the effect of recognition errors due to homophones can be reduced. Figure 16 shows some examples from retrieval outputs where the use of syllable-based units helps. From the recognizer vocabulary, we also found that there are 8,788 words with homophones. Among these words with homophones, there is an average of 2.34 words per pronunciation. Therefore, slight performance gain can be obtained by reducing the effect of errors due to these homophones.

Query:

Word : 廉 政 公署 繼續 調查 一 宗 . . .

Char2 : 廉政 政公 公署 署繼 繼續 續調 調查 . . .

Syl2 : . . . *gung_cyu* *cyu_gai* *gai_zuk* . . .

Document:

word : . . . 共 處 繼續 調查 . . .

Char2 : . . . 共處 處繼 繼續 續調 調查 . . .

Syl2 : . . . *gung_cyu* *cyu_gai* *gai_zuk* . . .

(a)

Query:

Word : . . . 當 見 習 售 貨 員

Char2 : . . . 當見 見習 習售 售貨 貨員

Syl2 : . . . *sau_fo* *fo_jyun* . . .

Document:

Word : . . . 交 收 貨 源 . . .

Char2 : . . . 交收 收貨 貨源 . . .

Syl2 : . . . *sau_fo* *fo_jyun* . . .

(b)

Figure 16 Comparison between syllable-based and character-based indexing units shows that syllable-based units are robust to errors due to homophones. In (a), the boxed characters come from a word but they are mis-recognized as other homophone characters in the document. Matching indexing units can only be found on syllable bigram scale as underlined. In (b), the boxed characters actually come from an OOV word and they are mis-recognized as other homophone characters. This kind of OOV word is prone to errors in recognition. By using syllable bigrams, all of the units from the OOV word can be recovered (in the form of subword bigrams) and contribute to the retrieval.

5.1.3 Comparison of SDR to textual retrieval results

When results of retrieval from textual summaries are compared to those from LVCSR outputs, we observe the following. Since there are recognition errors in transcriptions of spoken documents, retrieval from LVCSR outputs shows lower performance than retrieval from textual summaries. For retrieval from textual summaries, character-based indexing units perform better than syllable-based units. This is because textual summaries are free from recognition errors. Character-based units are also more discriminative than syllable-based units because there are homophones. However, the use of syllable-based units achieves better performance than character-based units in SDR. This marginal improvement for using syllable-based units is brought about from the robustness of syllable-based units to recognition errors due to homophones.

Table 6 shows the relative performance improvement of using subword bigrams over words. It is found that retrieval from LVCSR outputs achieves greater improvement (9%) than retrieval from textual summaries (1.7%). This phenomenon is observed for both syllable-based and character-based indexing units. The improved gain is obtained from the robustness of the subword bigrams to recognition errors.

Table 6 Relative gain in retrieval performance when subword bigrams are used instead of words. Both syllable- and character-based units are used for retrievals from textual summaries and LVCSR outputs.

	Bigrams	**Words**	**Relative gain**
		Textual summaries	
Syllable	0.834	0.819	**1.83%**
Character	0.839	0.826	**1.57%**
		LVCSR outputs	
Syllable	0.630	0.575	**9.56%**
Character	0.619	0.570	**8.60%**

5.2 Retrieval with multi-scale fusion

Multi-scale fusion experiments are carried out among different indexing scales, including syllable bigrams, character bigrams, and words. Words and subword bigrams are chosen for two reasons: 1) subword bigrams are found to achieve satisfactory retrieval performance for their robustness; 2) words are lexically-oriented and therefore contain information that is more specific. We attempt to combine the advantages of different indexing scales by data fusion approaches in our multi-scale fusion experiments. For pre-retrieval multi-scale fusion experiments, composite indexing scales are fused to form new representations in a larger space *before retrieval*. The same fusion process is applied to both the queries and documents and then retrieval is performed using the fused queries and documents as in (5) and (6). For post-retrieval multi-scale fusion experiments, fusion is carried out *after retrievals* and retrievals are carried out on the composite scales separately. Retrieval scores from these scales are then linearly combined as shown in (10) and documents in the collection are re-ranked according to the combined scores. For both pre-retrieval and post-retrieval multi-scale fusion approaches, retrieval experiments are performed using optimized weights for composite scales.

5.2.1 Multi-scale retrieval results

In Figure 17, results for the Cantonese SDR task using multi-scale fusion show that multi-scale fusion brings improvements to the retrieval performance. The best performance is obtained by fusion between syllable bigrams and words for both pre-retrieval fusion (AIR=0.641) and post-retrieval fusion (AIR=0.642). However, there is no significant improvement when character bigrams are fused with syllable bigrams.

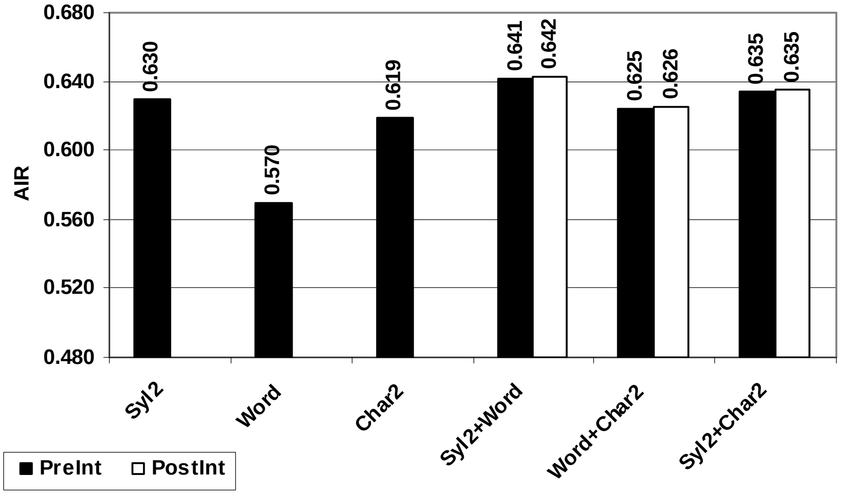


Figure 17 Retrieval results from multi-scale fusion of different indexing scales derived from LVCSR outputs. “Syl2” refers to syllable bigram scale, “Char2” means character bigram scale and “Word” means word scale.

5.2.2 Improvement by multi-scale fusion

Our multi-scale SDR experiments show that fusion of subword bigrams and words brings improvements in retrieval performance. It is believed that words and subword n-grams are complementary because words contain lexical information and subword n-grams are robust to OOV. By combining the specificity of word scale units together with the error robustness of subword scale units, improvement in retrieval performance is achieved.

As mentioned previously, the use of subword bigrams is beneficial since most words are two characters long and therefore these words can be covered by subword bigrams. The major gain by fusion of subword bigrams with words is the additional robustness to OOV words. However, fusing character bigrams with words does not achieve the same level of improvement as fusing syllable bigrams with words in SDR. This is because character bigrams are not robust to recognition errors due to homophones as with syllable bigrams. To illustrate this point, an example is extracted from the experimental results and shown in Figure 18 with detailed explanation provided in the caption.

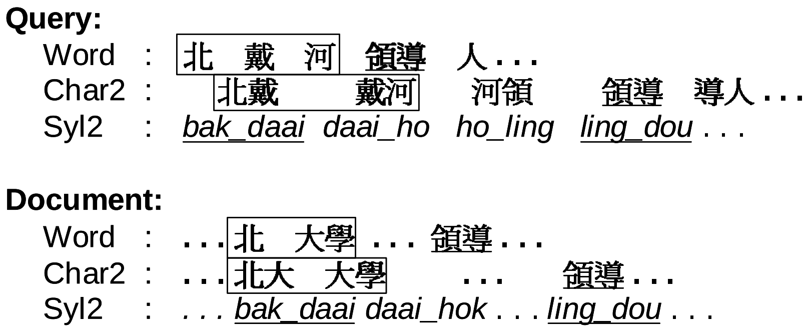

Query:
Word : 北 戴 河 領導 人...
Char2 : 北戴 戴河 河領 領導 導人...
Syl2 : bak_daai daai_ho ho_ling ling_dou . . .
Document:
Word : ...北 大學 ... 領導...
Char2 : ...北大 大學 ... 領導...
Syl2 : . . . bak_daai daai_hok . . . ling_dou . . .

Figure 18 Recognition errors due to homophones are overcome by retrieving using syllable bigrams in multi-scale SDR. The boxed word is an OOV word and mis-recognized by the recognizer as words with similar pronunciations. Retrieving with syllable bigrams can partially recover this OOV word by introducing an additional correct syllable bigram. Performance can further be improved by fusing with the word units for SDR to take advantage of the discriminating power of words. “Syl2” refers to syllable bigram scale, “Char2” means character bigram scale and “Word” means word scale.

Our retrieval results also show that fusing subword bigrams achieves little improvement. It can be explained by the fact that the two subword bigram scales carry similar information and neither of these scales have the specificity of words. As shown in Table 7, the two subword bigrams indexing scales give many equal ranking to documents when retrievals are performed separately. This means that the information captured by these indexing scales are similar and the room for retrieval performance improvement is small. On the other hand, higher performance gain can be obtained from fusing words and subword bigrams. This is because words and subword bigrams are complementary in the captured information. This can be verified from the smaller number of equally ranked documents between words and subword bigrams when compared to the pair of subword scales (see Table 7). As a result, multi-scale fusion can combine these scales to achieve performance improvements.

Table 7 The number of equally ranked documents between different pairs of indexing scales. “Syl2” refers to syllable bigram scale, “Char2” means character bigram scale and “Word” means word scale.

Textual summaries			**LVCSR outputs**		
Word	Char2	1377	Word	Char2	850
Syl2	Char2	1607	Syl2	Char2	1131
Syl2	Word	1354	Syl2	Word	841

5.3 *Consistency of performance improvement from multi-scale fusion*

In this section, we will experimentally demonstrate the consistency of retrieval performance improvement from multi-scale fusion. In multi-scale fusion, weights for the composite scales have to be optimized for achieving the best retrieval performance from fusion. The consistency of these scale weightings across different data collections is an important issue. It is because this will determine whether pre-optimized scale weightings could achieve performance improvement when applied to other data collections. In addition, these experiments also show that improvement to retrieval performance is achievable for spoken document collections of different sizes.

Multi-scale fusion of different indexing scales for improving retrieval performance relies on properly configured weightings for composite scales. In our multi-scale fusion experiments, these weights are optimized using the document collection. However, it is not always possible to optimize scale weightings in practice. Documents in the collection are usually changing continuously. One solution to this problem is to optimize the weights with a smaller subset of documents and apply the optimized weightings to the complete document collection. In order to verify that optimizing scale weightings this way is reliable and portable, we have divided the 1800-document collection into two mutually exclusive subsets of 1200 and 600 documents that are non-overlapping in time spanned. Scale weightings for these three sets of documents are searched empirically over the range from 0 to 1 (in steps of 0.1) in order to locate the best values. Figure 19 shows the search results for both pre-retrieval fusion and post-retrieval fusion.

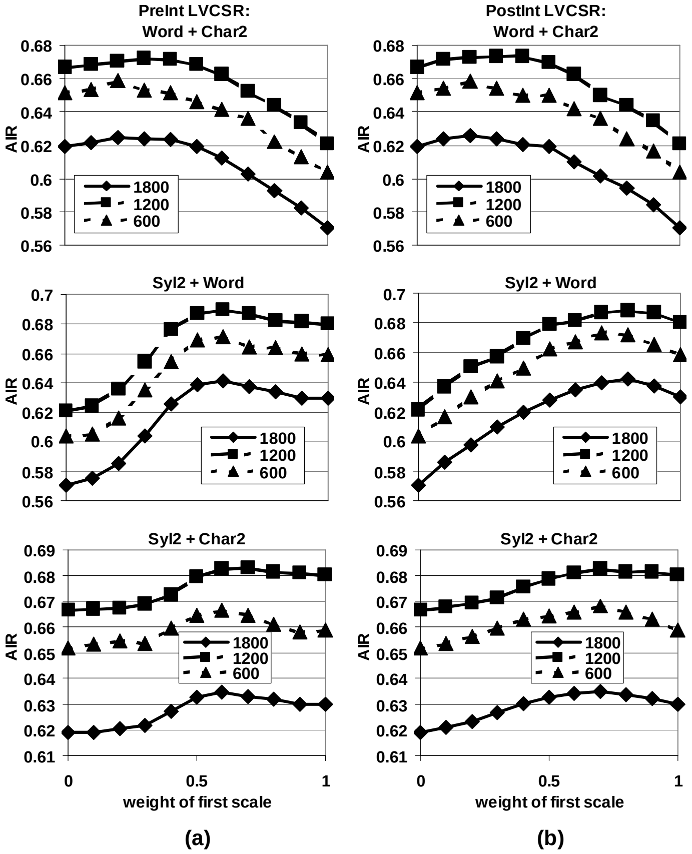


Figure 19 Illustration of the consistency of optimal weightings for composite scales in multi-scale fusion by exhaustive search over different subsets of document collections: the 1800-document full set, the 1200-document subset and the 600-document subset. The sums of the weights for composite scales are constrained to be 1. These search results show that optimal weightings are consistent across different subsets of documents for both (a) pre-retrieval (PreInt) and (b) post-retrieval (PostInt) fusion.

It can be seen that the optimal weights are consistent across the three sets of documents for both pre-retrieval and post-retrieval fusion. This consistency has two important implications. First, consistency across the 600-document subset and the 1800-document full set implies that by optimizing the scale weightings of a subset of the full collection, the optimized values are also “optimal” for the full collection. In addition, consistency across the 600-document subset and the 1200-document subset implies that optimized scale weightings in a small set of documents are also applicable to another *non-overlapping* document collection. This means that the optimized scale weightings are portable across different document collections (at least from the same source as in our case). For practical implementations, optimal scale weightings can be estimated from some training data. The optimization process may also be performed periodically to reduce the possible time dependency of scale weightings.

Another important implication of the consistent performance improvement across the three sets of documents is the applicability of multi-scale fusion. Even though the experimental collection is not large, performance improvements are obtainable across different subsets of the document collection. From the *non-overlapping* 600-document subset, 1200-document subset, to the full 1800-document collection, similar trends in the retrieval are observed as shown in Figure 19. These consistent trends mean that performance improvements are possible for document collections of different sizes.

6 Conclusions

In this work, we have presented the first investigation in spoken document retrieval (SDR) for Cantonese broadcast news. We have also compiled the first collection of Cantonese broadcast news for spoken document retrieval. Based on this collection, we have investigated the use of indexing units on word and subword scales for SDR. We found that subword bigrams achieve better retrieval performance over words due to the robustness to ambiguities in word segmentation and out-of-vocabulary problems. In addition, subword bigrams are also more robust to recognition errors than words. The use of subword bigrams as indexing units in Chinese SDR is beneficial and greater relative improvement (~9.56%) is achieved over words. To further improve the retrieval performance, we have adopted multi-scale retrieval by fusing both words and subwords in two different approaches: pre-retrieval fusion and post-retrieval fusion. Experiments have demonstrated that further improvements in retrieval performance (~1.90%) over using subword bigrams alone are achievable by fusing subword bigrams with words using both of fusion approaches. Our experimental results demonstrate that multi-scale retrieval by fusion of words and subword n-grams can achieve consistent improvements in SDR performance.

7 Acknowledgements

We would like to thank Television Broadcast (Hong Kong) Limited for permitting us to use their Internet news broadcast as our audio-video data archive in the experiments. We would also like to thank P.Y. Hui and Y.C. Li for their assistance.

8 Reference

Bartell, B. T., Cottrell, G. W., and Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval*, vol. 1, pp. 173–181.

Chen, B., Wang, H. M., and Lee, L. S. (2000a). Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 1771–1774.

Chen, B., Wang, H. M., and Lee, L. S. (2000b). Retrieval of mandarin broadcast news using spoken queries. In *Proceedings of the 6th International Conference on Spoken Language Processing*, vol. 1, pp. 520–523.

Chen, B., Wang, H. M., and Lee, L. S. (2001a). An HMM/n-gram-based linguistic processing approach for Mandarin spoken document retrieval. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, vol. 2, pp. 1045–1048.

Chen, B., Wang, H. M., and Lee, L. S. (2001b). Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, vol. 1, pp. 299–302.

Chen, B., Wang, H. M., and Lee, L. S. (2002). Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese. *IEEE Transactions on speech and audio processing*, 10:303-313.

Chien, L. F. (1995). 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*, pp. 112–120.

Chien, L. F. (1999). Pat-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. *Information Processing and Management*, 35:501–521.

Choi, J. Choi, Hindle, D., Hirschberg, J., Magrin-Chagnolleau , I., Nakatani , C., Pereira , F., Singhal , A., and Whittaker , S. (1998). Scan - speech content based audio navigator: a systems overview. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pp. 2867–2870.

Clarkson, P. R. and Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 2707–2710.

Compaq Research (2000). The 1st internet site for content-based indexing of streaming spoken audio. Technical report, Compaq Research, Cambridge MA. White paper: [online] *http://www.speechbot.com*.

Fox, E. A., Koushik, M. P., Shaw, J., and Modlin, R. (1992). Combining evidence from multiple searches. In *Proceedings of the 1st Text REtrieval Conference*, pp. 319–328.

Fox, E. A. and Shaw J. (1993). Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference*, pp. 243–252.

Foote, J. T., Young, S. J., Jones, G. J. F., and Jones, S. K. (1997). Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. *Computer, Speech and Language*, 11:207–224.

Hauptmann, A. G. and Witbrock, M. J. (1997). Informedia: News-on-demand multimedia information acquisition and retrieval. In *Intelligent Multimedia Information Retrieval*, pp. 213–239.

Johnson, S. E., Jourlin, P., Jones, K.J., and Woodland, P.C. (2001). Information retrieval from unsegmented broadcast news audio. *International Journal of Speech Technology*, 4:251–268.

Jones, G. J. F., Foote, J. T., Jones, K. S., and Young, S. J. (1996a). Retrieving spoken documents by combining multiple index sources. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 30–38.

Jones, S. K. J., G. J. F., Foote, J. T., and Young, S. J. (1996b). Experiments on spoken document retrieval. *Information Processing and Management*, 32:399–417.

Kwok, K. L.(1997). Comparing representations in Chinese information retrieval. In *Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 34–41.

Li, Y. C., Lo, W. K., Meng, H. M., and Ching, P. C. (2000). Query expansion using phonetic confusions for Chinese spoken document retrieval. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, pp. 89–93.

Lo, W. K., Meng, H. M., and Ching, P. C. (2000). Sub-syllabic acoustic modeling across Chinese dialects. In *Proceedings of the 2nd International Symposium on Chinese Spoken Language Processing*, pp. 97–100.

Lo, W. K., Schone, P., and Meng, H. M. (2001). Multi-scale retrieval in MEI: an English-Chinese translingual speech retrieval system. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, vol. 2, pp. 1303–1306.

Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R., and Srivastava, A. (2000). Speech and language technologies for audio indexing and retrieval. *Proceedings of IEEE*, 88:1338–1353.

Meng, H. M., Lo, W. K., Li, Y. C., and Ching, P. C. (1999). A study on the use of syllables for Chinese spoken document retrieval. Technical report, Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong, Hong Kong, China.

Meng, H. M., Lo, W. K., Li, Y. C., and Ching, P. C. (2000a). Multi-scale audio indexing for Chinese spoken document retrieval. In *Proceedings of the 6th International Conference on Spoken Language Processing*, vol. IV, pp. 101–104.

Meng, H. M., Chen, B., Grams, E., Khudanpur, S., Lo, W. K., Levow, G. A., Oard, D., Schone, P., Tang, K., Wang, H. M., and Wang, J. Q. (2000b). Mandarin-English information (MEI): Investigating translingual speech retrieval. Technical report, Johns Hopkins University, Baltimore, USA, 2000. Final report : [online] *http://www.clsp.jhu.edu/ws*2000/*final reports/mei*.

Meng, H. M. and Hui, P. Y. (2001). Spoken document retrieval for the languages in Hong Kong. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*, pp. 201–204.

Meng, H. M., Chen, B., Khudanpur, S., Levow, G. A., Lo, W. K. , Oard, D., Schone, P., Tang, K., Wang, H. M., and Wang, J. Q. (2001a). Mandarin-English information (MEI): Investigating translingual speech retrieval. In *Proceedings of the 2001 Human Language Technology Conference*.

Meng, H. M., Lo, W. K., Chen, B., and Tang, K. (2001b). Generating phonetic cognates to handle named entities in English-Chinese cross-language spoken document retrieval. In *Proceedings of the 2001 Automatic Speech Recognition and Understanding Workshop*.

Merlino, A. and Maybury, M. (1999). *An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News*. Cambridge: MIT Press.

Nie, J. Y. and Brisebois, M. (1996). On Chinese text retrieval. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 225–233.

Ng, K. (1998). Towards robust methods for spoken document retrieval. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pp. 939–942.

Ng, K. and Zue, V. W. (1998). Phonetic recognition for spoken document retrieval. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 325–328.

Ng, K. (2000a). Information fusion for spoken document retrieval. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 2405–2408.

Ng, K. (2000b). Subword-based approaches for spoken document retrieval. *Speech Communication*, 32:157–186.

Ng, K. (2000c). Towards an integrated approach for spoken document retrieval. In *Proceedings of the 6th International Conference on Spoken Language Processing*, vol. 1, pp. 672–675.

Nie, J. Y. and Ren, F. (1999). Chinese information retrieval: using characters or words? *Information Processing and Management*, 35:399–417.

Schauble, P. (1997). *Multimedia Information Retrieval - Content-based Information Retrieval from Large Text and Audio Databases*. Boston:Kluwer Academic.

Tuerk, A., Johnson, S. E., Jourlin, P., Jones, K. S., and Woodland, P. C. (2001). The Cambridge university multimedia document retrieval demo system. *International Journal of Speech Technology*, 4:241–250.

Thong, J. M., Goddeau, D., Litvinova, A., Logan, B., Moreno, P., and Swain, M. (2000). Speechbot: a speech recognition based audio indexing system for the web. In *Proceedings for the 2000 RIAO*.

Wang, H. M. (2000). Experiments in syllable-based retrieval of broadcast news speech in mandarin Chinese. *Speech Communication*, 32:49–60.

Wang, H. M., Meng, H. M., Schone, P., Chen, B., and Lo, W. K. (2001). Multi-scale audio indexing for translingual spoken document retrieval. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 605–608.

Wechsler, M. (2000). New approaches to spoken document retrieval. *Information Retrieval*, 3:173–188.

Whittaker, S., Hirschberg, J., Choi, J., Hindle, D., Pereira, F., and Singhal, A. (1999). Scan: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 26–33.

[i] Traditional Chinese characters are used in Hong Kong, Macau and Taiwan.

[ii] Large collections of Chinese documents have been made available through Text Retrieval Conference (TREC), Topic Detection and Tracking (TDT) projects, Linguistic Data Consortium (LDC), etc.

[iii] To cater for Internet broadcast, the clips are heavily compressed to 8.5 kbps using a CELP-based codec. Therefore, the extracted audio channels do not have high studio quality.