Multi-Scale Audio Indexing for Chinese Spoken Document Retrieval

Helen M. Meng*, W. K. Lo**, Yuk Chi Li*, and P. C. Ching**
*Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
**Digital Signal Processing Laboratory,
Department of Electronic Engineering,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong, China
{hmmeng,ycli@se.cuhk.edu.hk, wklo,pcching@ee.cuhk.edu.hk}

ABSTRACT

The advent of the information age has brought massive digital libraries of multimedia content. This development creates a high demand for information indexing and retrieval technologies, and the capability of browsing through audio archives is much desired. This paper reports on our initial attempt in the use of syllable units for Chinese spoken document retrieval. Our experiments are based on 1801 news stories from local television broadcasts in Cantonese, a monosyllabic Chinese dialect with a rich tonal structure. Results show that indexing with overlapping bi-syllables (tonal syllables) *mapped from text* delivers the reference retrieval performance at average inverse rank (AIR)=0.830. Retrieval based on overlapping bi-syllables (base syllables*) recognized from audio* achieved an AIR of 0.460.

I. INTRODUCTION

The advent of the information age has brought massive digital libraries of multimedia content. This development creates a high demand for information indexing and retrieval technologies, which enable the user to access personally relevant content on-demand. The capability of browsing through audio archives is desirable, and the technology is also applicable to video browsing, where we can index the video based on its audio track.

Previous work in this area includes the Informedia Project from CMU [1], which indexed audio tracks of news broadcast with by means of large-vocabulary speech recognition. The system has demonstrated multilingual capability in handling English as well as Serbo-Croatian. Ng and Zue [2] have studied the use of different subword units for audio indexing (e.g. broad classes, phone multigrams and syllables), as well as the use of robust retrieval methods; and experiments are based on English radio news broadcasts. Chien and Wang [3] have worked on spoken access of Mandarin Chinese spoken documents, which is based on syllable recognition and syllable string search. An informative overview of audio information retrieval can be found in [4]. Much work has also been done under the TREC[1] and TDT[2] programs sponsored by DARPA and NIST in the US.

This paper reports on our initial study in using syllables for spoken document retrieval. Our work is based on local Hong Kong television news broadcasts in Cantonese – the predominant Chinese dialect used in Hong Kong,[3] Macau, South China and many overseas Chinese communities. Cantonese is monosyllabic with a rich tonal structure – there are approximately 1,600 distinct tonal syllables with six lexical tones.

The syllable unit seems very desirable for Chinese spoken document retrieval, by virtue of the monosyllabic nature of the language and its dialectal variations. Hence syllables can fully characterize the language and provide high phonological coverage for spoken documents. This should enhance recall. When compared to the use of LVCSR or keyword spotting for audio indexing, subword indexing can circumvent the problem of incomplete keyword / vocabulary sets, as pointed out in [2]. The tradeoff, however, is a potential loss of discriminating power due to the exclusion of lexical information. A possible remedy may be the incorporation of *N*-gram sequential constraints.

We find several interesting research issues related to the use of subword (syllable-based) indexing for Chinese spoken document retrieval. They include:

(i) The use of tone information – we can compare retrieval results between indexing with *base syllables* (tone excluded) and indexing with *tonal syllables* (tone included). The acoustic correlates of tone, i.e. fundamental frequency and duration, are influenced heavily by prosodic context. This creates a challenge for highly accurate tone recognition. However, tone information should help discrimination and enhance retrieval performance.

(ii) The incorporation of sequential constraints in syllable-based audio indexing, and the contribution of such constraints towards retrieval performance. For instance, one may consider the use of bi-syllables, tri-syllables, or multi-syllables segments that are derived from lexical units.[4]

(iii) The effect of imperfect syllable recognition on retrieval performance.

This paper reports on a series of experiments designed to study the above issues.

II. SPOKEN DOCUMENTS FROM A VIDEO ARCHIVE

Our spoken documents are derived from a video archive of local television news broadcasts in Cantonese. We have collected a total of 1801 news stories within the six-month period June 1997 to February 1998. Each news story is a video

[1] http://trec.nist.gov/
[2] http://www.nist.gov/TDT
[3] Hong Kong's populace is trilingual, and speaks Cantonese, Mandarin as well as English.

[4] A Chinese word may consist of one or several characters. Each character is pronounced as a tonal syllable. Different characters may map to the same tonal syllable, and alternate pronunciations exist for some characters.

clip, conveniently packaged as a single RealMedia file, accompanied by a textual summary with a title.[5] Each video clip usually begins with an anchor presenting the gist of the story, but the opening may sometimes be a brief conversation between two anchors discussing the story. The opening is generally followed by a live report in the field, live coverage of the event, an interview, etc. Each accompanying textual summary is brief, and it is not a transcription of the audio track. Figure 1 shows an example of a textual summary.

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立會動議，要求將有關條文的有效期再延長多兩年。

Figure 1. Example of a textual summary accompanying the video clip of a news story. The summary title is underlined.

As mentioned, we have 1801 news stories in total. The textual summaries average 150 Chinese characters in length, and vary between 140 to 1700 characters. The video clips average 1.5 minutes in duration, and vary between 11 seconds to 25 minutes.

For the purpose of spoken document retrieval, we set aside the titles of the textual descriptions as queries. The rest of the textual description is treated as the textual document, with its corresponding audio document.

For the purpose of speech recognition development, we randomly extracted 2.25 hours of audio data to be orthographically transcribed by hand. This corresponds to the audio tracks of 157 news stories (totaling 1.75 hours) for training our recognizer, and 77 news stories (totaling 0.5 hours) for our development test set. The phonetic inventory for transcription follows the LSHK standard [5], and pronunciations are extracted from our own Cantonese pronunciation lexicons: CUPDICT and CULEX (CUPDICT has 10,346 entries and CULEX has 41,469 entries) [6].

III. THE RETRIEVAL METHOD

We use the vector-space model in SMART [7] for spoken document retrieval. A benchmark is provided by the use of textual queries to retrieve textual documents. Both queries and documents are first tokenized into Chinese words using a maximum-matching algorithm and the lexicon CULEX. The tokenized Chinese words become our indexing terms for retrieval. We have augmented the SMART to process Chinese characters for this purpose.

We first simulated spoken query retrieval of spoken documents by mapping the textual queries and documents into their syllable pronunciations, for both tonal and base syllables. The mapping also references CULEX. Indexing terms may thus be derived from the monosyllables, overlapping bi-syllables or overlapping tri-syllables. Overlap may be important to avoid segmentation errors. We also include "skipped bi-syllables"[6] to capture abbreviation terms as introduced in [8].

[5]The textual summary is primarily in Chinese, but occasionally may contain some numbers and English proper nouns.

[6] Skipped bi-syllables are formed from three consecutive syllables by skipping the middle one.

Additionally, we ran our speech recognizer on each audio track, to produce a syllable transcription of the spoken document. Using the same indexing procedure as described above, we can retrieve the spoken documents based on syllable pronunciations of the queries.

We adopted the following term weighing strategies for retrieval:

- For term *i* in document *d*:

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

where $tf_d[i]$ is the frequency of term *i* in document *d*

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

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

where *tf[i]* is the frequency of term *i* in query *q*

N is the total number of documents, and n_i is the number of documents with term *i*. The 0.5 in the above equations augments the relative *tf[i]* value. The similarity *S(q, d)* between a query *q* and document *d* is measured by the inner product, to form the basis of retrieval:

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

IV. SPEECH RECOGNITION

IV.1 Seed Models from Adapted Data

We have previously developed a Cantonese syllable recognizer, trained with phonetically-rich, continuous speech from the CUSENT corpus[9]. The corpus is recorded from a sound-proof room with a microphone. Therefore we need to adapt our corpus for indexing the audio tracks of television news broadcasts in RealAudio format. For this purpose, we downsampled CUSENT from 16kHz to 8kHz, and encoded the data with the RealAudio Encoder at 8.5kbps. We will refer to this adapted corpus as RA-CUSENT. RA-CUSENT is subsequently used for training seed acoustic models based on sub-syllable structures: Initials (I) and Finals (F) of the base syllables [10]. Syllable recognition is achieved by using a syllable dictionary of Initials and Finals.

Our acoustic models are CDHMM with mixture Gaussians, trained from MFCCs augmented with the first derivatives as well as an energy coefficient (26 parameters per input vector). We have trained context-dependent models of various mixtures (4, 8, and 16) – right-context-dependent Initial-Final model (BI_IF) and two-sided context-dependent models (TRI_IF). A accuracy-speed tradeoff suggests that we should use the BI_IF acoustic model with 16 Gaussian mixtures.

IV.2 Model Retraining

Our seed models (BI_IF, 16 mixtures) were trained from RA-CUSENT. They are *further retrained* using a small amount of hand-transcribed audio tracks. 2.25 hours of news stories are blindly segmented into 20-second fragments and transcribed by hand. 1.75 hours are used for retraining our acoustic models. The remaining 0.5 hour is used as our development test set, on which we obtained a syllable error rate of 73.75% using a syllable bigram. Recognition performance is low, as there is

great acoustic mismatch between RA-CUSENT and the audio tracks of our news broadcasts:

- CUSENT contains speech recorded with a microphone speech in a sound-proof room. The audio tracks contain broadcast speech recorded from the studio, in the field, over the telephone, etc.
- CUSENT contains Cantonese read speech, while the audio tracks contain more spontaneous speech. The tracks may also contain speech in languages other than Cantonese.
- The audio tracks also contain many non-speech sounds from live coverage of events.

We spot-checked two randomly selected news stories to examine the performance of our retrained recognizer. Results are shown in Table 1.

	Anchor	Reporter	Interviewee
Story 1	35.7%	34.0%	11.0%
Story 2	45.1%	35.1%	0% (very noisy)

Table 1. Syllable recognition accuracies of the retrained recognizer on two news stories. Performance measurements are displayed for different acoustic conditions

IV.3 Audio Indexing

The retrained acoustic models are used with a syllable bigram for indexing the audio tracks. Thus far we have processed 1801 news stories that contain approximately 54.5 hours of audio. The recognized base syllable strings from a single-pass Viterbi are used to index the news stories for retrieval.

V. EXPERIMENTS

V.1 Setup and Observations

Our experimental corpus does not provide readily available queries or relevance judgements for our study. Hence we formulated a *known-item retrieval* task using the following setup: The title from each textual summary is extracted and used as a query in retrieval. The remaining portion of the textual summary is treated as the "relevant" *textual* document. The corresponding audio track is treated as the "relevant" *audio* document. There are 1801 queries in total. Our task is to retrieve the relevant document for each query, and we adopted the *average inverse rank* as our evaluation metric:

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

where N equals to total number of news stories, $rank_i$ is the rank of the relevant document in the retrieved list for query i.

We have conducted a series of experiments described as follows (please refer to Figure 2):

1. The textual queries and documents were tokenized using CULEX (refer to the bars in Figure 2 with shading labeled TEXT). The retrieval based on Chinese words is shown as the TEXT bar labeled seg on the x-axis of Figure 2. The words are then broken down to individual characters to form character N-grams (N=1, 2 or 3) to be used as index terms (refer to the TEXT bars labeled mono, bi, tri on the x-axis in Figure 2). In addition, character bigrams augmented with skipped bigrams are also used (labeled as bi-skippedbi).
2. The textual queries and documents segmented above are mapped to tonal syllables for retrieval (refer to the bars in Figure 2 with shading labeled Tsyl(tonal)). The mapping references CUPDICT. Retrieval runs are then performed using the syllable unigrams (mono), bigrams (bi), trigrams (tri) and skipped bigrams (bi-skipbi).
3. Same as (2), except that *base* syllables are used (refer to bars in Figure 2 with shading labeled Tsyl).
4. We ran our syllable recognizer on the audio documents using a syllable bigram (refer to bars in Figure 2 with shading labeled Rsyl). The recognition outputs are then used to form syllable N-grams and skipped bigrams, which are used as index terms in our retrieval runs.

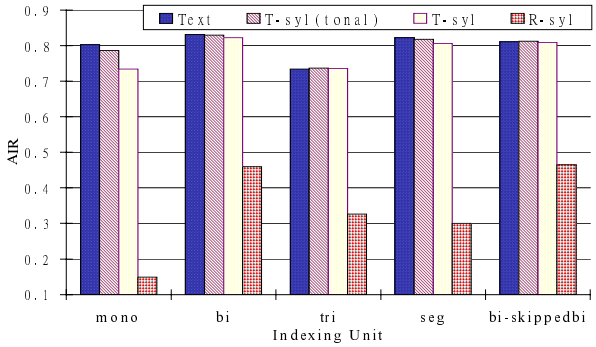

Figure 2. Retrieval results (AIR) of text, transcribed tonal syllable, transcribed base syllable and recognized base syllable for various kinds of index unit. This is based on a known-item retrieval task for 1801 queries.

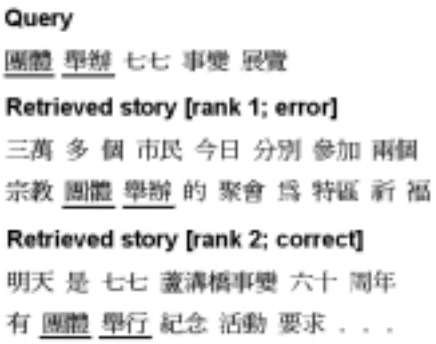

Figure 3. Example of a retrieval error due to alternate wording between the query and the document. The correct story is ranked second in the retrieved list of documents.

V.2 Analysis of Results

Noteworthy findings based on results in Figure 2 include:

(i) The Text Retrieval Benchmark: Word-based retrieval gave AIR=0.822. Errors are often due to the use of abbreviations in the queries (while the full form is used in the documents); and the use of different wordings between queries and documents. An example is shown in Figure 3. The word 舉行 was used instead of the word 舉辦 in targeted story and thus worsen retrieval result. Character bigrams gave an AIR of 0.832 possibly by better coverage on tokenization ambiguities.

(ii) Tonal Syllables versus Base Syllables: As observed in Figure 2, the use of tonal syllables has a small but consistent advantage over base syllables to produce an AIR gain of up to 0.053 (for monosyllables). This suggests the importance of accurate tone recognition in audio indexing. Providing tone information adds discriminating power in the set of index terms, especially in the case of monosyllables.

(iii) Benefits of Sequential Constraints: The incorporation of sequential constraint using bi-syllables produces an AIR gain over monosyllables – AIR gains range from 0.029 to 0.311. However indexing with bi-syllables also outperforms indexing with tri-syllables. Investigation on the query tokenization with CULEX shows that over 45% of the query terms are two-character words. Only 2.32% of the query terms are three-character words. This suggests that the syllable trigram context may be too specific for effective retrieval in the current corpus.
(iv) Use of Lexical Knowledge: All overlapping bi-grams (bars labeled bi in Figure 2, including characters or syllables) fared better in retrieval than tokenized terms (bars labeled seg). We surmise that this is due to out-of-vocabulary (OOV) words, e.g. many proper names and abbreviations are absent from CULEX. The tokenizer tends to segment an OOV into a sequence of single-character words. (See example in Table 2). In these cases the correct index terms are reduced to single-character word sequences and sequential constraints are lost.

As we migrate from indexing with Chinese words \character *N*-grams to indexing with their syllable pronunciations, there is a reduction in lexical knowledge. This is due to the homophone effect, where a Chinese character maps to 1.15 base syllables and 1.21 tonal syllables on average. Character *N*-grams fared better in retrieval than the corresponding syllable *N*-grams. For the bars labeled seg in Figure 2, AIR=0.822 when we index with Chinese words. This is reduced to 0.819 for tonal syllables and AIR=0.806 for base syllables. However, the lexical knowledge preserved in the syllable-based word pronunciations drastically reduces the number of index terms, when compared with the use of overlapping syllable *N*-grams. (See Table 3).

在 六月 三十 日 同 七月 一日 晚上 維 港 將 舉行 煙 花

Table 2. Word tokenization errors are mainly caused by out-of-vocabulary words present in our news corpus but absent from CULEX. In this table, the underlined examples should be two-character Chinese words. The first example is the short form for Victoria Harbor, and is wrongly segmented as two single-character words. The second example is a two-character word which means "fireworks". It is segmented into the single-character words *(smoke)* and *(flower)*.

	Seg. Text	**Mono Syl.**	**Bi-Syl.**	**Tri-Syl.**	**Seg. Syl.**
No Tone	6632	514	29,870	82,382	4,896
With Tone	6632	1140	40,342	85,055	5,630

Table 3. Number of distinct index terms in the various cases. (**Seg. Text** refers to tokenized Chinese words and Seg. Syl. refers to their word pronunciations. **Mono- Bi-** and **Tri-Syl** refer to the overlapping syllable *N*-grams (N=1,2 and 3).

(v) Speech Recognition: Retrieval was also conducted with speech recognition outputs that are base syllable sequences. As shown in Figure 2, retrieval performance degraded for all kinds of index terms. AIR for overlapping syllable bigrams is 0.460. This is caused by recognition errors during indexing. The transcribed portion of our corpus shows that on average the transcribed recognition output is four times as long as the textual summary (by an average syllable count of 304.7 versus 70.8). The extra information may ameliorate the effects of our errorful audio indexing for retrieval. However, the issue warrants deeper investigation.

VI. CONCLUSIONS & FUTURE WORK

This paper reports on our initial attempt in spoken document retrieval using the audio tracks of local television news broadcasts in Cantonese. We studied the use various syllable-based units for audio indexing, which include base syllables and tonal syllables as monosyllables, overlapping syllable bigrams and trigrams, as well as syllable bigrams with skipped syllable bigrams. We formulated a known-item retrieval task based on a video corpus of 1801 news stories, which have audio tracks accompanied by textual summaries. The text is mapped into syllables by referencing the CUPDICT and CULEX lexicons. Many words and terms in the news corpus are absent from our lexicons, and these affected our retrieval results based on text. Indexing with *mapped* overlapping bi-syllables (with tone) delivers a reference retrieval performance with AIR=0.830. Retrieval based on *recognized* overlapping bi-syllables (no tone) gave AIR=0.46. In the near future, we plan to incorporate query expansion strategies for retrieval.

VII. ACKNOWLEDGMENTS

CUSENT, CUPDICT and CULEX are part of the CU Corpora, a Cantonese corpora designed and collected with support from the Hong Kong Government's Industrial Support Fund. We thank the Television Broadcast (HK) Limited for providing the broadcast news stories in this project.

VIII. REFERENCES

[1] Hauptman, A. et al., "Multilingual Informedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages," Proceedings of the DARPA Workshop on Broadcast News Understanding Systems, Lansdowne, VA, February, 1998.
[2] Ng, K., "Towards Robust Methods for Spoken Document Retrieval," Proceedings of ICSLP, Sydney, 1998.
[3] Chien, L. F. and H. M. Wang, "Exploration of Spoken Access for Chinese Text and Speech Information Retrieval," Proceedings of the International Symposium on Signal Processing and Intelligent Systems, 1999.
[4] Foote, J. "An Overview of Audio Information Retrieval." ACM-Springer Multimedia Systems., 1997.
[5] LSHK, "Cantonese Transcription Table," the Linguistic Society of Hong Kong, 1997
[6] http://dsp.ee.cuhk.edu.hk/speech.
[7] Salton, G. and M. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, NY 1983.
[8] Chen, B., H. M. Wang and L. S. Lee, "Retrieval of Broadcast News Speech in Mandarin Chinese Collected in Taiwan using Syllable-Level Statistical Characteristcs," Proceedings of ICASSP, Budapest, 2000.
[9] Lo, W. K., T. Lee and P. C. Ching, "Development of Cantonese Spoken Language Corpora for Speech Applications", Proceedings of ISCSLP, Singapore, 1998.
[10] Wong, Y. W. et al., "Acoustic Modeling and Language Modeling for Cantonese LVCSR", Proceedings of Eurospeech, Budapest,1999.