RETRIEVAL OF BROADCAST NEWS SPEECH IN MANDARIN CHINESE COLLECTED IN TAIWAN USING SYLLABLE-LEVEL STATISTICAL CHARACTERISTICS

Berlin Chen[1,2], Hsin-min Wang[1], and Lin-shan Lee[1,2]

[1]Institute of Information Science, Academia Sinica,
[2] Dept. of Computer Science & Information Engineering, National Taiwan University,
Taipei, Taiwan, Republic of China

ABSTRACT

Spoken document retrieval has been extensively studied in recent years because of its high potential in various applications in the near future. Considering the monosyllabic structure of Chinese language, a whole class of indexing features for retrieval of spoken documents in Mandarin Chinese using syllable-level statistical characteristics has been studied, and very encouraging experimental results on retrieval of broadcast news speech collected in Taiwan were obtained. This paper reports some interesting initial results and findings obtained in this research.

1. INTRODUCTION

The network technologies and the Internet activities have created a completely new information era. Intelligent and efficient information retrieval techniques providing Internet users with easy access to spoken documents, such as broadcast radio and television programs, become highly desired and have been extensively studied in recent years [1-6]. At the same time, the DARPA Hub-4 contest that began in 1995 has been evaluating the technologies of use of large-vocabulary continuous speech recognition (LVCSR) to transcribe audio recordings of broadcast news with many evaluation results reported [7] so far. Regardless of all these developments, automatic recognition and efficient retrieval of broadcast radio and television news speech remains to be a very challenging research topic because of the wide variety of speaking styles and acoustic conditions [8-10], and the various different problems in spoken document retrieval.

There have been several different approaches developed for spoken document retrieval (SDR) in recent years. Word–based retrieval approaches have been very popular and successful, although with the potential problems of either having to know the query words in advance, or requiring a large enough lexicon to cover the growing dynamic contents of the diverse broadcast news [3]. Some other researchers proposed the concept of subword-based approaches, also with some other potential problems to be solved yet. A syllable-based indexing approach for retrieval of Chinese speech information using speech queries developed earlier at Taipei [6] belonged to the latter case. In this paper, the recent results of a new phase of research on the retrieval of real world Mandarin broadcast news speech collected in Taiwan area are reported. A whole class of indexing features based on the syllable-level statistical characteristic features has been studied. This is based on the various considerations on the structural nature of Chinese language. Very encouraging results are obtained and included.

2. CONSIDERATIONS OF USING SYLLABLE-LEVEL STATISTICAL CHARACTERISTICS

In Chinese language, because each of the large number of characters (at least 10,000 are commonly used) is pronounced as a syllable, and is a morpheme with its own meaning, new words are very easily generated everyday by combining few characters. For example, the combination of the character “電(electricity) and “腦(brain)” gives a new word “電腦(computer)”, and the combination of the characters “股(stock) ”, “市(market)”, “長(long)”, and “紅(red)” gives a new word “股市長紅 (stock price remains high for long)” in business news. In many cases the meaning of these words have to do with the meaning of the component characters. Examples of such new words also include many proper nouns such as personal names and organization names which are simply arbitrary combinations of a few characters, as well as many domain specific terms just as the examples mentioned above. Many of such words are very often the right key in retrieval functions, because they usually carry the core information, or characterize the subject topic. But in most cases these important words for retrieval purposes are simply not included in any lexicon. It is therefore believed that the out-of-vocabulary problem is especially important for Chinese information retrieval, and this is a very important reason why the syllable-level statistical characteristics makes great sense in the problem here. In other words, the syllables represent characters with meaning, and in the retrieval process they do not have to be decoded into words which may not exist in the lexicon.

Actually, the syllable-level information makes great sense for retrieval of Chinese information due to the more general monosyllabic structure of the language. Although there exist more than 10,000 commonly used Chinese characters, a nice feature of Chinese language is that all Chinese characters are monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1,345. So a syllable is usually shared by many homonym characters with completely different meanings. Each Chinese word is then composed of from one to several characters (or syllables), thus the combination of these 1,345 syllables actually gives almost unlimited number of Chinese words. In other words, each syllable may stand for many different characters with different meanings, while the combination of several specific syllables very often gives only very few, if not unique, homonym polysyllabic words. As a result, comparing the input query and the documents to be retrieved based on the segments of several syllables may provide very good degree of similarity between them, as can be found in the experiments to be presented later on in this paper.

In fact, there exist other important reasons to use syllable-level information. Because almost every Chinese character is a morpheme with its own meaning, thus very often plays quite independent linguistic roles. As a result, the construction of Chinese words from characters is very often quite flexible. One example phenomenon is that in many cases different words describing the same or similar concepts can be constructed by slightly different characters, e.g., both "中華文化(Chinese culture)" and "中國文化(Chinese culture)" means the same, but the second characters used in these two words are different. Another example phenomenon is that a longer word can be arbitrarily abbreviated into shorter words, e.g., "國家科學委員會(National Science Council)" can be abbreviated into "國科會", which includes only the first, the third and the last characters. An intelligent retrieval system needs to be able to handle such wording flexibilities, such that when the input queries include some words in one form, the desired audio recordings can be retrieved even if they include the corresponding words in other different forms. The comparison between the speech queries and the audio recordings directly at the syllable-level does provide such flexibilities to some extent, since the "words" are never actually constructed during the retrieving processes, while the different forms of words describing the same or similar concepts very often do have some syllables in common.

3. BROADCAST NEWS DATABASE

In this paper, the radio news was recorded using a wizard FM radio connected to a PC, and digitized at a sampling rate of 16kHz with 16bit resolution. The whole broadcast news database was collected from December 1998 to July 1999. The training data consists of 453 stories (about 4.0 hours of speech materials), and was collected from Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT), all located at Taipei. All recordings were manually transcribed and segmented into stories and sentences. In this research, only the speech parts of the anchor speakers were used in the experiments. The testing data to be retrieved consists of 757 recordings (about 10.2 hours of speech materials), and was collected from Broadcasting Corporation of China (BCC). Each recording was a short news abstract (about 50 seconds) produced by an anchor speaker, and contained several news items. Some recordings in the testing data contain background music. The testing data was also manually transcribed but was not segmented into sentences. So, the experiments here roughly correspond to the partitioned evaluation (PE) with the baseline broadcast condition (F0) in the 1996 Hub-4 test [11]. Table 1 shows the average speaking rates of the radio news from respective sources. It can be found that, on average, there are 5.8 characters per second in the testing radio news speech (BCC), while the average speaking rate for the training speech is 5.4 characters per second.

Data Type	Radio News Speech					
	Testing	Training				
Source	BCC	PRS	UFO	VOH	VOT	Average
Average Speaking Rates (characters/sec)	5.8	4.9	6.2	4.8	5.7	5.4

Table 1: Speaking rates of broadcast news speech

4. SYLLABLE RECOGNITION

4.1 Acoustic Processing and Language Modeling

Each frame of the speech data is represented by a 39 dimensional feature vector, which consists of 12 MFCCs and log energy, and their first and second differences. Utterance-based cepstral mean subtraction (CMS) is applied to all the training and testing materials. The acoustic units chosen for syllable recognition here are 112 context-dependent INITIAL's and 38 context-independent FINAL's, specially considering the phonetic structure of Mandarin syllables. Here INITIAL is the initial consonant of the syllable and FINAL is the vowel (or diphthong) part but including optional medial or nasal ending. Each INITIAL was represented by a HMM with 3 states while each FINAL with 4 states. The Gaussian mixture number per state ranged from 2 to 16, depending on the amount of training data. The silence model was a 1-state HMM with 32 Gaussian mixtures trained with the non-speech segments. In addition, a breath model, which was also 1-state HMM with 32 Gaussian mixtures, was specially trained to model the breath effect frequently occurs in broadcast news speech.

The syllable N-gram language models (including Bi-gram and Tri-gram) were applied in the speech recognizer to improve the recognition accuracy. They were trained by a text corpus consisting of 50 million Chinese characters collected from Central News Agency (CNA) in 1999 from January to July. Note that both the text corpus and the broadcast news database were collected almost in the same time frame. Word segmentation [12] and phonetic spelling were performed for the training materials using a lexicon consisting of around 85,000 frequently used Chinese words [13].

4.2 Syllable Recognition and Initial Tests

The simplest syllable recognizer performs only free syllable decoding without any grammar constraints. In a more complicated syllable recognizer a two-pass search strategy is used. In the first pass, Viterbi search was performed based on the acoustic models and the syllable Bi-gram language model, and the score at every time index were stored. In the second pass, a backward time-asynchronous A* tree search [14] generates the best syllable sequence based on the heuristic scores obtained from the first pass search and the syllable Tri-gram language model. Based on the state likelihood scores calculated in the first pass search and the syllable boundaries of the best syllable sequence, the syllable recognizer further performs the Viterbi search on each utterance segment which may include a syllable and outputs several most possible syllable candidates, and a syllable lattice can thus be easily constructed.

Some initial experiments were performed with results shown in Table 2. In the second row, the test using the acoustic models trained by a different set of read speech training database gives the recognition rate for the radio news speech only 25.8%, is obviously very poor. However, when the acoustic models trained with the radio news speech training database mentioned above were used, the recognition rate of the radio news speech can be immediately improved to 55.7%, as can be found in the third row of Table 2. Finally, the syllable *N*-gram language models were applied, and the results are shown in last two rows in the table.

	Syllable Recognition Rates
Models Trained by Read Speech	25.8%
Models Trained by Radio News Speech	55.7%
Models Trained by Radio News Speech Plus LM	64.9%
Models Trained by Radio News Speech Plus LM (Top 10)	89.2%

Table 2: Syllable recognition results for radio news speech

The syllable recognition rate can be further improved to 64.9% with a Top10 inclusion rate of 89.2%. Such syllable accuracy gives the syllable lattices to be used in the following retrieval task. The use of more complicated acoustic processing, training or adaptation schemes are currently under progress.

5. RETRIEVAL WITH SYLLABLE-LEVEL STATISTICAL CHARACTERISTICS

5.1 Vector Space Models

Vector space models widely used in many text information retrieval systems were used here, in which each document or query is represented by a feature vector V ,

$$V = (w_1 \times f(t_1) \times idf(t_1),...,w_k \times f(t_K) \times idf(t_K)) \quad (1)$$

where w_i , $f(t_i)$ and $idf(t_i)$ are respectively the weight, score and the inverse document frequency (IDF) of the indexing term t_i , while K is the total number of distinct indexing terms. The weight w_i is different for different classes of indexing terms as explained below. The Cosine measure widely used in text information retrieval was used to estimate the similarity between a document and a query [15].

5.2 Syllable-level Indexing Terms

Here a whole class of syllable-level indexing terms were used, including syllable segments with length N ($S(N)$, N=1,2,3,4,5) and syllable pairs separated by n syllables ($P(n)$, n=1,2,3,4). Considering a syllable sequence of 10 syllables $S_1 S_2 S_3 \dots S_{10}$, examples of the former are listed on the upper half of Table 3, while examples of the latter on the lower half of Table 3. For example, syllable segments of length N=3 includes such segments as (S_1 S_2 S_3), (S_2 S_3 S_4), etc., while syllable pairs separated by n=1 syllables includes such pairs as (S_1 S_3), (S_2 S_4), etc. Consider the structural features of the Chinese language, combinations of these indexing terms makes good sense for retrieval process. For example, as mentioned previously, each syllable represents some characters with meaning, and very often words with similar or relevant concepts have some syllables in common. Therefore syllable segments with length N=1 makes sense in retrieval process. However, because each syllable is also shared by many homonym characters, such syllable segments with length N=1 also causes ambiguity. Therefore it has to be combined with other indexing terms. On the other hand, more than 90% of most frequently used Chinese words are bi-syllabic, so the syllable segments with length N=2 definitely carry a plurality of linguistic information, and make great sense to be used as an important indexing term. Similarity, if longer syllable segments with N=3 are matched between a document and the query, very often this bring very important information for retrieval purposes. On the other hand, because of the very flexible wording structure in Chinese language as described previously, syllable pairs separated by n syllables are helpful in retrieval. For example, when the word "國家科學委員會 (National Science Council)" is abbreviated by including only the first, third and the last characters. Syllable pairs separated by n syllables becomes useful. Furthermore, because substitution, insertion and deletion errors always happen during the syllable recognition process, such indexing terms as syllable pairs separated by n syllables are also helpful in these problems.

Syllable Segments	Examples
$S(N)$, N=1	(s_1) (s_2) ...(s_{10})
$S(N)$, N=2	$(s_1\ s_2)$ $(s_2\ s_3)$…$(s_9\ s_{10})$
$S(N)$, N=3	$(s_1\ s_2\ s_3)$ $(s_2\ s_3\ s_4)$…$(s_8\ s_9\ s_{10})$;
$S(N)$, N=4	$(s_1\ s_2\ s_3\ s_4)$ $(s_2\ s_3\ s_4\ s_5)$…$(s_7\ s_8\ s_9\ s_{10})$
$S(N)$, N=5	$(s_1\ s_2\ s_3\ s_4\ s_5)$ $(s_2\ s_3\ s_4\ s_5\ s_6)$…$(s_6\ s_7\ s_8\ s_9\ s_{10})$
Syllable Pair Separated by n Syllables	Examples
$P(n)$, n=1	$(s_1\ s_3)$ $(s_2\ s_4)$ …$(s_8\ s_{10})$
$P(n)$, n=2	$(s_1\ s_4)$ $(s_2\ s_5)$ …$(s_7\ s_{10})$
$P(n)$, n=3	$(s_1\ s_5)$ $(s_2\ s_6)$ …$(s_6\ s_{10})$
$P(n)$, n=4	$(s_1\ s_6)$ $(s_2\ s_7)$ …$(s_5\ s_{10})$

Table 3: Various indexing methods for an example syllable string $S_1 S_2 S_3 \dots S_{10}$.

In each indexing term t_i discussed here, the acoustic recognition scores of the component syllables can be combined together as the term score, $f(t_i)$ in equation (1). Here the acoustic recognition score of every syllable candidate is first normalized with respect to the duration of the observed speech segment, and then transformed into a range between 0 and 1 by a Sigmoid function.

6. EXPERIMENTAL RESULTS FOR THE RETRIEVAL TASK

The 757 news recordings collected from BCC as explained previously were used in the following SDR experiments. Each recording contains about 50 seconds of news abstract pronounced by an anchor speaker. A set of 40 query terms and the corresponding relevant news recordings were manually created to support the retrieval experiments. Each query has on average 23.3 relevant documents among the 757 in the database, with the exact number ranging from 1 to 75. Two (one male and one female) speakers were asked to pronounce the 40 query terms respectively. The average syllable recognition rate for these spoken queries is 81.5%. At first glance this SDR task seems easy since the retrieval target contains only anchors' speech. However, in fact almost all query terms appear only once in each of their relevant documents, and the query terms are often very short. On average, each query term contains roughly only 4 characters (or syllables). As a result, this SDR task is in fact very difficult.

In order to obtain the upper bound text-based retrieval performance as references, the text transcriptions of both the queries and documents were also automatically converted into

	S(N), *N*=1~2	*S(N)*, *N*=1~3	*S(N)*, *N*=1~4	*S(N)*, *N*=1~5	*S(N)*, *N*=1~3 *P(n)*, *n*=1	*S(N)*, *N*=1~3 *P(n)*, *n*=1~2	*S(N)*, *N*=1~3 *P(n)*, *n*=1~3	*S(N)*, *N*=1~3 *P(n)*, *n*=1~4	*S(N)*, *N*=1~5 *P(n)*, *n*=1~3
TQ/TD	91.7%	96.3%	96.1%	96.1%	96.5%	96.7%	97.1%	97.1%	97.1%
SQ/TD	73.2%	77.6%	77.2%	77.2%	79.3%	79.2%	79.4%	79.0%	79.4%
TQ/SD	52.2%	55.4%	55.4%	55.4%	55.4%	57.3%	57.9%(69.9%*)	55.3%	57.9%
SQ/SD	42.5%	45.4%	45.4%	45.6%	46.8%	48.3%	48.4%(55.5%*)	47.0%	49.5%

Table 4: Retrieval results with respect to various combinations of syllable-level indexing terms
(*: Syllable lattice with verification and optimal $\overline{m}$)

syllable strings and performed IR experiments. These exactly correct query and document transcriptions are denoted as TQ (**T**ext **Q**ueries) and TD (**T**ext **D**ocuments). The erroneous syllable strings/lattices of queries and documents obtained from speech recognition are denoted as SQ (**S**poken **Q**ueries) and SD (**S**poken **D**ocuments) correspondingly. This gives a total of 4 categories of retrieval (TQ/TD, SQ/ TD, TQ/SD, SQ/SD) experiments and the results in terms of the widely used *non-interpolated average precision* [16] using different combinations of the syllable-level indexing terms are summarized in Table 4, in which *S(N)*, *N*=1~2 means both syllable segments with *N*=1 and 2 are used and so on. For all the four categories of retrieval, it can be found from the left part of the table that the retrieval performance is significantly improved when the syllable segment indexing is extended from *N*=2 to *N*=3, but the improvements when N is further extended to 4 or 5 become relatively insignificant. These results are in parallel with those obtained previously [3], using phone-level (a phone is a smaller acoustic unit than a syllable) indexing approach for English. When the syllable pairs separated by n syllables, *P(n)*, were used in addition as indexing terms as in the right parts of Table 4, it can be found that in general the retrieval performance can be further improved. However, the information of syllable pairs separated by too many syllables does not help. The best feature vector is thus *S(N)*, *N*=1~3 plus *P(n)*, *n*=1~3 in general.

The above results are based on the Top1 syllable recognition results only, whose accuracy is 64.9%. In fact, the Top10 recognition candidates with inclusion rate of 89.2% can be used here. However, when the number of syllable candidates is increased from 1 to m, $1 \le m \le 10$, the number of syllable segments *S(N)* or syllable pairs *P(n)* for *N*=*n*=2, for example, is increased from 1 to m^2. Although one of them may be exactly correct and provide the right information, the other $m^2 - 1$ all carry wrong syllables, and therefore inevitably increase the degree of ambiguity. The situation becomes even worse when *N* or *n* is larger. A syllable verification scheme was therefore used to filter out first the syllable candidates with lower scores by a threshold. The depth of the syllable lattice thus can be adjusted by the threshold value. In this way, a more accurate but compact syllable lattice can be obtained. Using the same indexing terms in the best cases of the above experiment, the average precision can be further improved from 48.4% to 55.5% for the SQ/SD case, and from 57.9% to 69.9% for the TQ/SD case, as also listed in Table 4. The average syllable candidate number $(\overline{m})$ is about 2.4 for the spoken documents while 1.3 for the spoken queries.

7. CONCLUSIONS

This paper presented only initial results of a project for long term research. The results are certainly not yet satisfactory. Further improvements including various directions of research are currently under progress. For example, to compare as well as integrate with the popular approach using character-level or word-level information will be definitely interesting.

8. REFERENCES

[1] D. A. James, "The Application of Classical Information Retrieval to Techniques to Spoken Documents," Ph.D. thesis, University of Cambridge, UK, 1995.

[2] K. Spärck Jones, G. J. F. Jones, J. T. Foote and S. J. Young, "Experiments on Spoken Document Retrieval," *Information Processing & Management*, 32(4), pp. 399-417, 1996.

[3] K. Ng and V. Zue, "Phonetic Recognition for Spoken Document Retrieval," In Proc. *ICSLP*, 1998.

[4] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition," Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich, 1998.

[5] CMU Informedia Digital Video Library project http://informedia.cs.cmu.edu/

[6] B. R. Bai, L. F. Chien, and L. S. Lee, "Very-Large-Vocabulary Mandarin Voice Message File Retrieval Using Speech Queries", In Proc. *ICSLP*, 1996.

[7] Proceedings of DARPA speech recognition workshop, http://www.itl.nist.gov/div894/894.01/proc/index.htm.

[8] T. Matsuoka, Y. Taguchi, K. Ohtsuki, S. Furui, and K. Shirai, "Toward Automatic Transcription of Japanese Broadcast News," In Proc. *EuroSpeech* 1997.

[9] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, and S. J. Young, "Experiments in Broadcast News transcription," In Proc. *ICASSP*, 1998.

[10] S.S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. A. Olsen, "Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News," In Proc. *ICASSP*, 1999.

[11] J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and Preparation of the 1996 Hub-4 Broadcast News benchmark Test Corpora," In Proc. *DARPA Speech Recognition Workshop*, 1997.

[12] K. J. Chen and S. H. Liu, "Word Identification for Mandarin Chinese Sentences," In Proc. *COLING*, 1992, pp. 101-107.

[13] CKIP group, "Analysis of Syntactic Categories for Chinese," *CKIP Technical Report*, no. 93-05, Institute of Information Science, Academia Sinica, Taipei, 1993.

[14] B. Chen, H. M. Wang, L. F. Chien, and L. S. Lee, "A*-Admissible Key-Phrase Spotting With Sub-Syllable Level Utterance Verification," In Proc. *ICSLP*, 1998.

[15] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, NY, 1983.

[16] D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)," 1995, Available at "http://trec.nist.gov/pubs/trec4/overview.ps".