Fast Spoken Term Detection Using Pre-retrieval Results of Syllable Bigrams

Hiroyuki Saito[†], Yoshiaki Itoh[†], Kazunori Kojima[†], Masaaki Ishigame[†], Kazuyo Tanaka[††], Shi-Wook Lee[†††]
[†]Iwate Prefectural University, Faculty of Software, Japan
[††]Tsukuba University, Japan
[†††]National Institute of Advanced Industrial Science and Technology, Japan
E-mail: y-itoh@iwate-pu.ac.jp

***Abstract*— We propose a method of the Spoken Term Detection (STD) based on a priori retrieval results in which plural syllables are used as query terms. In the proposed method, all N-syllable combinations such as syllable bigrams are searched for in spoken documents. In the first step of the method, the retrieval results are prepared a priori, where pre-retrieval results include candidates with scores matching those of each N-syllable sequence. Given a query, the syllable sequence of the query is divided into plural syllable sequences whose lengths are the same as those of the pre-retrieval results. In the second step, the candidate sections are filtered by using the scores of query's syllable combinations. This reduction in the number of candidate sections for detailed matching leads to a large reduction of the retrieval time. In the third step, these candidates sections are re-scored by performing detailed matching. Experimental results show that the proposed method reduces the retrieval time by 93% with a performance degradation of less than 2 points.**

I. INTRODUCTION

Spoken term detection (STD) is a key technology that is demanded to search for scenes of interest in a section of large scale multimedia data. Many STD studies have been presented recently at the Text Retrieval Conference (TREC) and the 9th Workshop of the National Institute of Informatics Test Collection for Information Retrieval Systems (NTCIR-9) [1-2]. In many STD systems, the recognition results of an automatic speech recognizer can be utilized if the query word is found in the recognizer's dictionary. However, the selected query word is often a special term that is not included in such dictionaries. STD systems must therefore be based on an open vocabulary, and use subword units such as monophones and triphones to deal with unknown words. Improvement of retrieval performance and reduction of retrieval time are then important problems in subword-based systems. The retrieval performance of unknown query words based on subword units is lower than that of known query words based on word units. Moreover, considerable time is needed to match all subword recognition results in spoken documents.

This paper proposes a fast STD method that uses pre-retrieval results of N syllables. The initial concept of this method was that all types of queries are retrieved and the results prepared a priori. However, this method fails with unknown query words. Instead of retrieving all query words, the proposed method retrieves all combinations of N-syllable sequences such as bigrams, which are then prepared a priori in the first step. The proposed method thus enables a reduction of the matching process by referring to the pre-retrieval results generated from the spoken documents. Given a query in Japanese, the syllable sequence of the query is automatically obtained because Japanese words are inherently composed of syllable sequences. In the second step of the method, the syllable sequence is divided into multiple sequences whose lengths are equal to those of the pre-retrieval results. In this step, the candidate sections are filtered by using the scores of the query's syllable combinations. This process reduces the number of candidates sections. In the third step, we perform detailed matching against these reduced candidate sections and rescore the candidate sections.
Although many studies have been conducted on subword-based STD systems and methods to increase their retrieval speeds [3-6], the majority of these methods have used an inverted index or a suffix array to construct an index of subword recognition results. Performing a robust approximate search on such an index is difficult considering the deletion and insertion errors. In contrast, the proposed method enables approximate searching by retrieving syllable bigrams in advance, and candidate sections can be selected without removing sections that contain errors since these are accounted for in the detailed matching.

In the present paper, the proposed STD system and the method of using N-syllable's pre-retrieval results are explained in detail. The performance of the proposed method is then evaluated through experiments using the open test collections of NTCIR-9 [2]. Finally, conclusions are presented.

II. PROPOSED METHOD

This section describes the proposed fast STD method that uses pre-retrieval results of N-syllables.

A. Outline of previous STD system

A process chart outlining our previous STD system is shown in Figure 1. Subword acoustic models, their associated language models, a subword distance matrix, and subword recognition results of spoken documents are all prepared a priori [7,8]. First, subword recognition is performed for all of the spoken documents and a subword sequence database is

prepared in advance (1). Subword language models such as subword bigrams are then employed. When a user inputs a text query, it is automatically converted into a subword sequence according to a set of conversion rules (2). The phonic sequence of a user-input query term to be pronounced is automatically obtained from the term's syllables. By using Continuous Dynamic Programming (CDP) algorithms, the system then retrieves the target section by comparing a query subword sequence with all subword sequences found in the spoken documents (3). We use the acoustic distance as a local distance measure, where the acoustic local distance represents the dissimilarity between two subwords and indicates the statistical distance between any two subword models. The system outputs multiple candidate sections that have a high degree of similarity to the query word. Each candidate section thus has a distance and a section number of spoken documents.

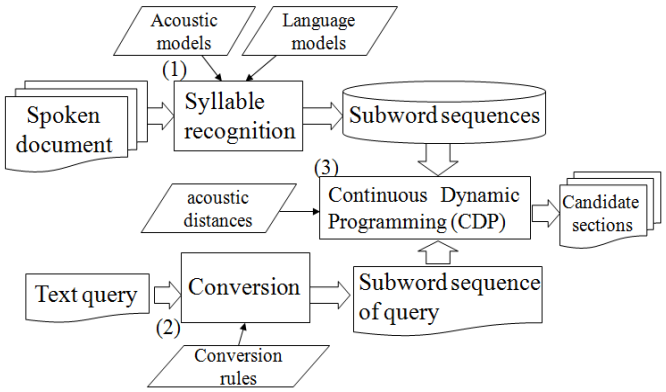

Fig 1. Outline of our previous STD system.

This system must perform matching between a query and all spoken documents because we use the acoustic distance as a local distance measure in CDP. Such matching leads to a long computation time that is linearly proportional to the total length of the spoken documents. Under the conditions of NTCIR-9 SpokenDoc STD subtask [2], our previous system takes 0.953 s to obtain all candidates for a query from spoken documents totaling about 44 h. Because the length of spoken documents in a larger subtask in NTCIR-9 was about 600 h and the retrieval time of the previous system was proportional to the spoken documents, the speed-up of the system was inevitable.

B. Introduction of pre-retrieval results of N-syllables

We can assume without loss of generality that any statement that includes a query word also includes all syllable sequences that make up the query. For example, if the query ABCD is composed of a sequence of four syllables A, B, C, and D and a statement includes the query, we can assume that the statement also includes both of the sequences ABC and BCD in the syllable trigram case. Thus, when performing retrieval using ABC and BCD as queries, the candidate sections are also thought to be candidates of the original query ABCD. The target sections can be limited by using the retrieval results of such sub-syllable sequences of the query. Therefore, pre-retrieval results are prepared a priori by retrieving all N-syllable combinations. When a query is given by a user, the syllable sequence of the query is divided into plural sub-syllable sequences whose lengths are equal to those of the pre-retrieval results. By conducting detail matching by CDP for only those sections included in the pre-retrieval results of the query's sub-syllable sequences, fast retrieval is realized.

The process of the proposed method is outlined in Figure 2, where the proposed method is composed of the following five procedures.

1. Subword recognition using syllables, triphones, and so on is performed for all spoken documents such that they are converted to subword sequences.
2. Pre-retrieval is conducted by using all N-syllable combinations as queries, and highly ranked candidate sections are retained for each N-syllable sequence. Here, let K denote the number of candidate sections.
3. When a text query is given by a user, the query is automatically converted into a syllable sequence, which is then divided into sub-syllable sequences by shifting an N-syllable window. When the length of a query is L, the number of sub-syllable sequences is L-N+1.
4. Pre-retrieval results are referred to for each sub-syllable sequence in the query, and all pre-retrieval results for the sub-syllable sequences are merged. These candidate sections are called the "first candidate sections".
5. Detailed matching by CDP is performed only for the first candidate sections, which are rescored and submitted to the user as "last candidate sections" according to their ranked order.

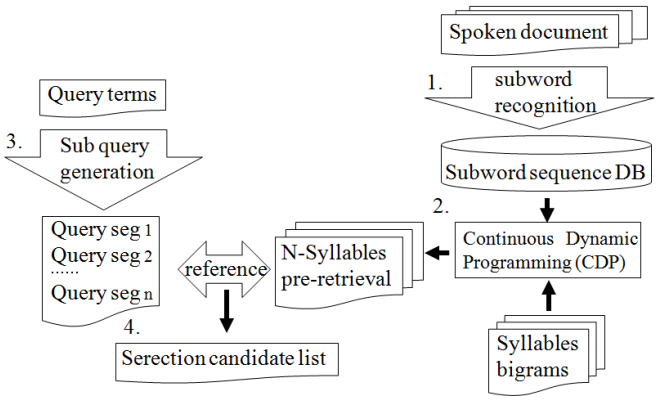

Fig 2. Proposed STD system using pre-retrieval results.

The first and second procedures are performed in advance and are termed the "first step" of the method. The third and fourth procedures are then termed the "second step", and the last procedure is termed the "third step". Various merging methods are available for the fourth procedure; however, we use the union of all pre-retrieval results in the present paper.

Although the second procedure states that candidate sections within the K candidates are retained, the number of candidate sections is actually controlled by the threshold of the CDP distance scores.

C. *Introduction of pre-retrieval results of N-syllables*

In preliminary experiments, we discovered that about 18 h are needed to construct pre-retrieval results of spoken documents totaling about 55 h by using syllable bigrams (N = 2). The number of Japanese syllables amounts to 261 and the computation time is 0.953 (s/query)×261^2 (query) ≈ 18 h. Furthermore, around 284 MB of memory is needed to store the index when K = 5000. For N = 3, 196 days and 72GB of memory are needed to construct and store the results. Therefore, we decide to do an experiment using a more realistic syllable bigrams.

III. EVALUATION EXPERIMENTS

A. *Experimental Conditions*

We used the open test collections of the NTCIR-9 workshop in 2011. The test collections, which were dry run sets in NTCIR-9, include 50 queries and 177 presentation speech data or spoken documents that are the so called "core" data of the Corpus of Spontaneous Japanese (CSJ) database. About 2,500 presentation speech data other than the core are also contained in the CSJ database for training acoustic models and language models. The feature extraction conditions used here for acoustic models are listed in Table 1. Julius version 4.1.5 was used as the syllable recognition decoder in experiments.

Table 1: Feature extraction conditions for acoustic models.

Sampling	16 kHz 16 bit
Feature Parameter	12-dim. MFCC 12-dim. Δ MFCC + Δenergy 12-dim. ΔΔMFCC + ΔΔenergy
Window Length	25 ms
Frame Shift	10 ms for monophone and triphone

We used three states and right-to-left Hidden Markov Models (HMM) as acoustic models of each triphone, and syllable bigrams and trigrams as language models. These models were trained by using the 2,500 presentation speech data sets mentioned above that are not included in the evaluation set. Because we distinguish between short and long vowels to construct a syllable, 261 Japanese syllables were used in the experiments, double the number of general Japanese syllables. We prepared syllable bigram pre-retrieval results by retrieving all syllables bigram combinations (261^2) in the spoken documents a priori. Candidate sections were then stored in hash memory.

We used mean average precision (MAP) as an evaluation measure of performance, which was also employed in NTCIR-9. Average precision (AP) for a query is obtained by averaging the precisions at every correct occurrence, and MAP is then the averaged AP of all queries. The processing time was obtained when using a personal computer with an Intel Core i7 2600 processor, 8 GB of memory, and a Linux operating system.

B. *Results for pre-retrieval results of syllable bigrams*

Figure 3 shows the retrieval performance (MAP (%); bar graphs) and retrieval time (s; line graph) for a query according to the number of candidates sections (the Top-K). "all" in the figure denotes the case of conducting CDP for all speech datasets. MAP at Top-5000 did not declined compared with that at "all", and the retrieval time was reduced from 0.953 to 0.404 s. The results indicate that each syllable bigram does not have to retain more than 5,000 candidate sections.

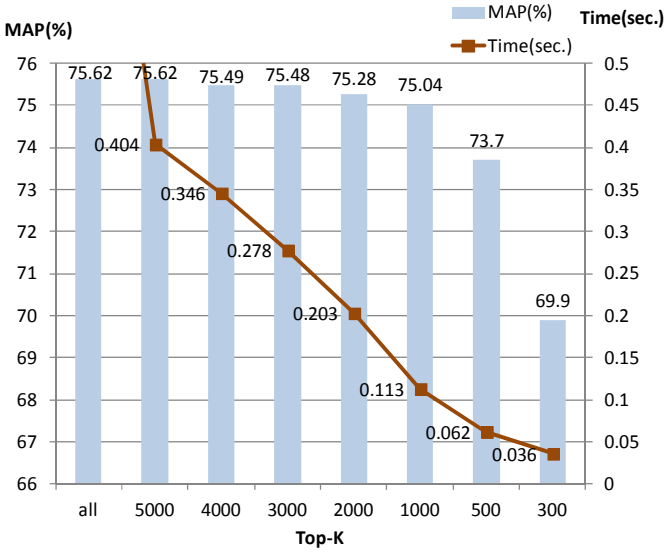

Fig 3. Performance of pre-retrieval results for syllable bigrams.

At Top-1000, the retrieval time was reduced by about 88% to 0.113 s with only a 0.58 point performance decline. At Top-500, the retrieval time was reduced by about 94% to 0.062 s with a 1.92 point performance decline. We have thus verified that the proposed method can reduce retrieval time while sustaining retrieval performance. Because the performance decline was 5.7 points at Top-300, we conducted the next experiment by using the Top-500 to Top-5000 candidate sections.

C. *Effect of filtering candidate sections by distance threshold*

Because the retrieval time is approximately proportional to the number of candidates sections, we now describe the effect of filtering these candidate sections by adding a distance threshold. Threshold values from 6.0 to 9.0 were tested for the Top-500 to Top-5000 candidate sections.

In Figure 4, the bar and line graphs show the retrieval performance (MAP; %) and retrieval time (s) for a query, respectively. The graphs corresponding to "no threshold" and "no Top-K" indicate the "all" case in Figure 3. Compared with the case without a threshold, the retrieval time was reduced by introducing the distance threshold. For example, by introducing a distance threshold 7.0 at Top-3,000, the performance deteriorates by 0.62 points (from 75.62% to 75.01%) and the retrieval time reduces by 42% (from 0.278 to 0.161 s). Among the four threshold values, performance deteriorates by less than 2.0 points at a threshold value of 7.0

or greater. Here, 7.0 corresponds to one-third of the maximum acoustic local distance between two triphone models.

The use of both the Top-K candidate sections and distance threshold filtering was shown to be effective in sustaining the retrieval performance when comparing two cases in Table 2.

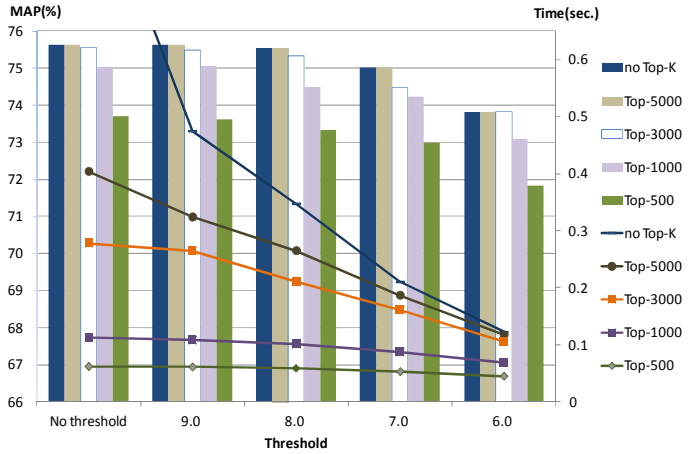

Fig 4. Performance of pre-retrieval result with distance threshold filtering

IV. CONCLUSIONS

This paper has proposed an STD method based on obtaining pre-retrieval results in advance by using syllable bigrams as query terms. All combinations of syllable bigrams are searched for in a set of spoken documents, and are prepared as pre-retrieval results. The candidate sections are then filtered by using the scores of pre-retrieval results for the query's syllable bigrams. The resulting reduction in the number of candidates sections for detailed matching then leads to fast retrieval. Results of experiments demonstrated that the proposed method could reduce the retrieval time by 66.0% (retrieval is 2.27-fold faster) without degrading retrieval performance. By adding a distance threshold to filter thecandidate section, the retrieval time could be reduced by 93.59% (15.6-fold faster), with less than a 2.0 points degradation of the retrieval performance.

Table 2: Effect of the use of both Top-*K* and threshold filtering

	MAP	Time reduction
No Top-*K* with 6.0 threshold	73.81%	87.19%
Top-1000 with 10.0 threshold	75.04%	88.24%

In future work, we plan to investigate better scoring method for using the pre-retrieval results than the simple union currently used. A method is also required to use such scores directly to rank the candidate sections without having to perform detailed matching. The performance when using syllable trigrams should also be evaluated.

ACKNOWLEDGMENT

This research is supported by Grand-in-Aid for Scientific Research (C) Project No. 20500096, KAKENHI of Japan Society for Promotion of Science.

REFERENCES

[1] Auzanne C., Garofolo J. S., Fiscus J. G., Fisher W. M., "Automatic Language Model Adaptation for Spoken Document Retrieval," B1, 2000TREC-9 SDR Track, 2000.

[2] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara and Tomoko Matsui, "Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop", 9th NTCIR Workshop, pp.223-235, 2011.

[3] Taisuke Kaneko, Tomoyosi Akiba, "Metric Subspace Indexing for Fast Spoken Term Detection", INTERSPEECH 2010.

[4] Joel Pinto Igor Szoke S.R.M. Prasanna, Hynek Hermansky, "Fast Approximate Spoken Term Detection from Sequence of Phonemes," IDIPA RESEARCH REPORT 2008.

[5] David R. H. Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Colthurst, Stephen A. Lowe, Richard M. Schwartz, Herbert Gish Thomas , "Rapid and Accurate Spoken Term Detection," INTERSPEECH 2007.

[6] Kouichi Katsurada, Shinta Sawada, Shigeki Teshima, Yurie Iribe and Tsuneo Nitta, "Evaluation of Fast Spoken Term Detection Using a Suffix Array", INTERSPEECH2011, pp.909-912 2011.

[7] Yoshiaki Itoh et al., "Two-stage Vocabulary-free Spoken Document Retrieval -Subword Identification and Re-recognition of the Identified Sections-", ICSLP, pp. 1161-1164, 2006.

[8] Yoshiaki Itoh et al., "Spoken Term Detection Results Using Plural Subword Models by Estimating Detection Performance for Each Query," INTERSPEECH, pp. 2117 -2120, 2011.