WEB Page Collection Using Automatic Document Segmentation for Spoken Document Retrieval

Hiromitsu Nishizaki*, Kiyotaka Sugimoto† and Yoshihiro Sekiguchi*
* Department of Research Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan
† Department of Education Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan
E-mail: {hnishi,sekiguti}@yamanashi.ac.jp Tel/Fax: +81-55-220-8361/8776

***Abstract*—In spoken document retrieval, the main factor affecting retrieval performance is speech recognition errors. Refining speech recognition technology can make improvement of speech recognition performance. However, if a query has out-of-vocabulary words, we cannot get the spoken documents related to the query. This paper describes spoken document retrieval using document expansion based on WEB whose contents are similar to the spoken documents retrieved. Most of spoken documents have some topics. Therefore, each spoken document is automatically divided into some segments depending on topic. And then, similar WEB pages to the spoken document can be collected using the query derived from the segment. The document expansion using WEB achieved improvement of the spoken document retrieval performance from 0.364 to 0.401 on interpolated 11-points average precition metric.**

I. Introduction

A rapidly increasing number of spoken documents, such as broadcast programs, spoken lectures, and recordings of meetings are archived, and some of them can be accessed through the Internet. The need to retrieve such spoken information has been growing, while an effective retrieval technique is definitely lacking at present; thus, the development of technology for retrieving such information is becoming increasingly important.

In the TREC Spoken Document Retrieval (SDR) track hosted by NIST and DARPA in the latter half of the 1990s, a number of studies of SDR were presented on the subject of English and Mandarin broadcast news documents.

A standard approach to spoken document retrieval is to automatically convert spoken documents into word sequences using a Large Vocabulary Continuous Speech Recognition (LVCSR) system. The transcribed word sequences can be directly matched against queries. In this approach, however, a serious problem arises when both the queries and the documents include out-of-vocabulary (OOV) words; matching against OOV words always fails, because the OOV word cannot be transcribed as a word. Many previous studies handled the OOV problem in spoken document retrieval. In most of them, spoken documents were not transcribed into word sequences, but into sequences of sub-words (such as phonemes and syllables) using sub-word recognizers [1], [2], [3], [4]. For example, K. Ng et al. [1] worked on the use of sub-word unit representation based on phonemes in spoken document retrieval. C. Ng et al. [2] reported, however, that better retrieval performance in terms of average precision was obtained using word-based indexing rather than phonemes/syllables based indexing. Furthermore, word based retrieval is necessarily faster than that based on phonemes/syllables.

This paper investigates the effectiveness of document expansion for the spoken lectures retrieval task by using WEB pages. There are two advantages for using document expansion using WEB on an SDR task. One of advantages is robust for speech recognition errors and OOV. The other is to supplement keywords that are not uttered in the spoken document, which a user wishes to look for. For solving the speech recognition errors and OOV problems, we prepared two kinds of indexes: One is made from transcriptions of spoken documents by an LVCSR system; the other is made from WEB pages related to the target spoken documents. In our approach, WEB pages are retrieved by a search engine that uses WEB search queries automatically composed from transcriptions of the target spoken documents.

A similar approach to our proposed method has been already reported. Singhal et al. [5] adopted document expansion for news speech retrieval by using a news text corpus related to the target news speech. This approach was easy, because speech recognition of news speech was very high, and it is easy to find text documents similar to the target news speech.

In this paper, we aim to retrieve Japanese spoken lecture, which are included in the Corpus of Spontaneous Japanese (CSJ) [6]. This is difficult task because speech recognition for these lecture, including filled pauses and areas of dis-fluency, is not easy, and the topics of the speeches are wide ranging. Therefore, it is hard to automatically look for the documents that are used in document expansion.

When document expansion techniques are used in the retrieval of these documents, WEB pages may be the most suitable, because the Internet has a wide variety of topics. Consequently, in this paper, we investigate the effectiveness of document expansion by using WEB pages on SDR.

The problem of document expansion using WEB for SDR is to how to collect WEB pages whoes contents are similar to the spoken documents. We performed WEB pages collection by human power, then, the SDR performance was drastically improved compared with the performance without the document expansion. However, it is hard to collect suitable WEB pages. So, we propose an WEB collection method using an automatic spoken document segmentation method [7]. Most of the spoken documents have some topics. Collecting WEB page for each topic may improvement the quarity of WEB pages.

Experimental results have shown that the our document expansion was effective for solving the recognition errors and OOV problems. The retrieval peroformance got inprovement of relative 3.8%. In particular, the retrieval performance for only the queries including OOV words has relatively improved by 49.3% compared with the baseline result. Furthermore, the document expansion with the document segmenatation got

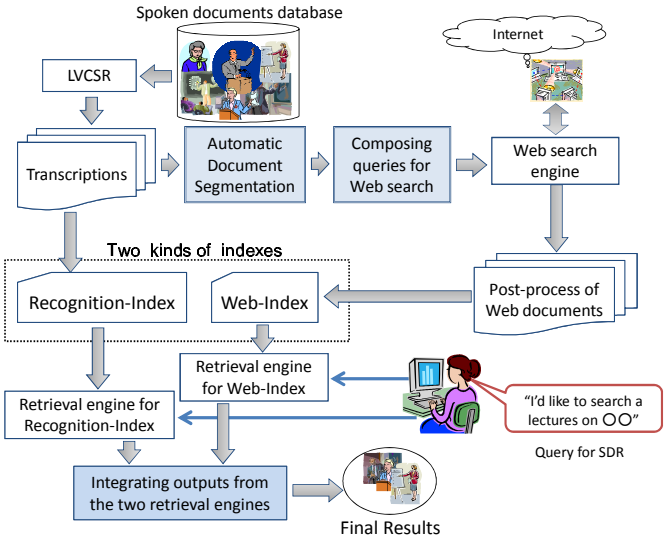

Fig. 1. *A framework of spoken document retrieval by using document expansion using WEB.*

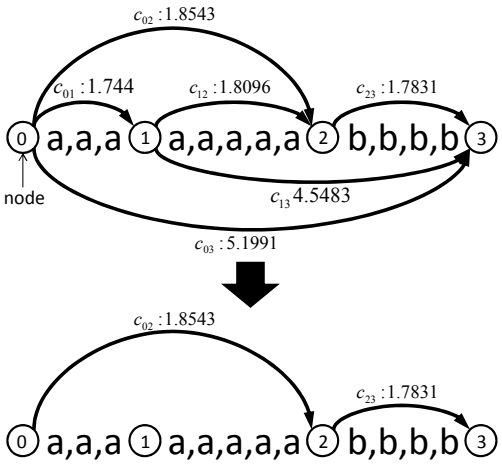

Fig. 2. *Example of automatic document segmentation.*

more better performance than the one without the expansion (relative 6.1% improvement).

II. Spoken Document Retrieval Using Document Expansion

Figure 1 shows the outline of our proposed technique.

First, spoken documents are automatically transcribed by an LVCSR system. The document index is built by removing stop words from the transcriptions. In this paper, we call it "Recognition-Index" (**RI**).

Next, the other index is made from WEB pages as follows:

1) Each the transcription is automatically divided into some segments depending on topic.
2) Queries for WEB searches are composed from the transcriptions of segmented spoken documents. For each spoken document, multiple queries, which depend on the number of segments of a spoken document, are prepared.
3) For each segment, an WEB search engine collects WEB pages from the Internet using the query. Most of the collected WEB pages may be related to the part of spoken document from which the search query is made.
4) Stop words are removed from the collected WEB pages. The WEB-based index, which we call the "WEB-Index" (**WI**), is made from them.

The performance of retrieval using the WI depends on how the WEB search query is composed. In addition, the quality of the query relies on speech recognition performance during transcription of the spoken document. However, the object of this study is to investigate whether the WI is effective in spoken document retrieval. Therefore, we adopt a very simple query composition method, as described in Section II-B.

In the spoken document retrieval process, two retrieval engines search the spoken documents, each using one of the indexes. A query is input to each retrieval engine; the final retrieval results are obtained by integrating their two outputs based on a retrieval score attached to each retrieval document.

A. Automatic Document Segmentation

To use WEB pages in spoken document retrieval, queries for an WEB search engine are composed by extracting keywords from transcriptions derived by recognizing spoken documents. However, the transcriptions include a number of words unsuitable for documents on a specific topic. Therefore, it is difficult to extract keywords that well-represent the topic of a spoken document.

To solve this problem, first, the spoken documents are automatically divided into some segments, then, queries are composed from each segment. Most of spoken documents have several kinds of topics. The segmentation of spoken documents makes WEB search queries to be better for collecting WEB pages.

We use the text segmentation algorithm proposed by Utiyama et al.[7] for spoken document segmentation. Although the algorithm was adopted to word sequences in [7], we extended the algorithm to accept sentence sequences.

First, we put nodes on the beginning of the document, the end of the document, and between all two successive sentences. The cost value c_{ij} for a segment S_{ij} between two arbitrary nodes i and j is defined as follows:

$$c_{ij} = \sum_{l=1}^{length} \log \frac{length + k}{tf_l} + penalty \times \log W \qquad (1)$$

where $length$ is the total number of words contained in a segment S_{ij} and k is the number of kinds of words in the whole document. tf_l means term frequency of word w_l occured in S_{ij} and W is the total number of words in the whole document. $penalty$ is a hyper-parameter that can control the number of segments. The dynamic programming (DP) scheme using the cost can indicate some segmentation nodes by minimizing the total DP costs.

Figure 2 shows an example of document segmentation. There are three sentences: S_{01} ="aaa", S_{12} ="$aaaaa$", and S_{23} ="$bbbb$" and the $penalty$ parameter is set to 1.0 in Figure 2. In this case, k and W are 2 and 12, respectively. For example, the cost of the segment S_{03} can be calculated by using Equation(1) and its value is $c_{03} = 5.1991$. Finally, the document is divided into two segments: S_{02} and S_{23} because the total cost is $c_{02} + c_{23} = 3.6374$, which is smaller than c_{03}.

B. Query Composition

The query composition method we adopted was very simple. Queries for a spoken document are composed of word-based N-grams that occur with very high frequent in the document's transcription. The procedure contains only five steps, as follows:

1) A transcription of a spoken document is automatically divided into M segments by the automatic segmentation scheme explained in Section II-A.
2) Word-based N-grams are extracted from M segmented transcription. An N-gram is denoted as a sequence of specific part of speech, namely, a noun and postpositional particle. The length of N is not limited.
3) Some of the N-grams do not exist in the real world because these are extracted from a transcription containing many speech recognition errors. These are filtered out by using the corpus of "WEB Japanese N-gram 1st edition," which contains N-grams made from WEB data collected by Google Japan, Inc. In this paper, we call the corpus "Google N-gram".
4) Stop words that are included in the N-grams are removed. Then, the five most frequent N-grams are extracted from each segment.
5) Finally, M N-gram query sets (each set has five N-grams) is used to search WEB pages. Up to maximum 50 WEB pages are collected for each spoken document. The number of WEB pages collected by a query from a segment depends on the number of sentences in the segment. If a segment has 50% of sentences in a document, the query from the segment can get 25 WEB pages.

C. Collecting WEB Documents and Making WEB-Index

For each query, WEB pages are collected by an WEB search engine using the query made from the transcription of the spoken document. We used "Yahoo! WEB search API" as the search engine in this study.

Stop words are removed from the collected WEB pages. WEB pages collected by a query set from a segment are integrated into a file. Each file corresponds to one specific segment. Finally, WI is made of them.

D. Spoken Document Retrieval Engine

We used the "Generic Engine for Transposable Association" (GETA) [8] as the spoken document retrieval engine. The GETA can realize fast computation of similarity between a query vector and document vectors.

In this research, word-unit indexes (RI and WI) needed to retrieve spoken documents are constructed from only the content words of transcriptions of spoken documents (RI) or WEB documents (WI) by removing stop words. Nouns, verbs, and adjectives are adopted as content words.

Queries for spoken document retrieval are sentences in the form of "List some World Heritage sites," for example, so morphological analysis is performed on the query to segment it into a word sequence. After removing stop words from the sequence, the query is input to the GETA engine.

The computation of similarity between a query vector and document vectors is done by the SMART method [9], which is based on cosine similarity, like TF-IDF and is available in the GETA. Each indexed word is weighted by the TF-IDF method, in which the TF value of each word is normalized on document length (number of words).

TABLE I
Recognition performance and OOV rates.

# of words	Corr. [%]	Acc. [%]	OOV [%]	# of OOV
17k	76.9	71.6	11.8	11

E. Integrating Outputs from Two Retrieval Engines

The final retrieval results are obtained by integrating the outputs from two retrieval engines. One retrieval engine uses RI, and the other uses WI. Each engine outputs a list of spoken documents, in order of similarity score.

Therefore, we can get the final retrieval results (a list of documents) by combining the two similarity scores from RI and WI. The final combined similarity score $sim(d)$ for a spoken document from the two engines is calculated as follows:

$$sim(d) = (1-\alpha) \times sim(d|r) + \alpha \times \max_{s \in S} sim(d_s|w) \quad (2)$$

where $sim(d|r)$ is the score from the RI engine. Suppose that $S = \{d_1, d_2, \cdots, d_M\}$ is a segment set when a document d is divided into M segments. So, $sim(d_s|w)$ is the score of segment s for document d from the WI engine.

In retrieval experiments, α is incremented until 1.0 from 0.0 by 0.1 point, because it is difficult to set the suitable value of α. However, we should set a unique value of α in a real spoken document retrieval system.

III. Retrieval Experiment

A. Test Collection for Spoken Document Retrieval

We used the test collection for evaluating spoken document retrieval [10], which consists of a set of textual queries and relevant segment lists, allowing retrieval from the CSJ [6]. The test collection has been developed by the Spoken Document Processing Working Group, which is part of a special interest group of the Information Processing Society of Japan.

The CSJ has 2702 lectures that are recorded at academic meetings and simulated meetings. In this paper, one lecture is denoted as one document. The 39 Japanese queries included in the test collection are prepared for retrieving lectures in the CSJ.

B. Speech Recognition for Spoken Documents

As described in Figure 1, the 2702 spoken documents in CSJ are transcribed by the LVCSR system, which is called "Julius" ver.4.1 for indexing.

Julius, an open source decoder for LVCSR, has a trigram language model (LM) and a triphone-based acoustic model. Julius used triphone-based HMMs trained from the large set speeches with the high quality headset microphone, which were sampled at 16kHz and 16bits. Feature vectors consist of 38 dimensions: 12 dimensional Mel-frequency cepstrum coefficients (MFCCs) , the cepstrum difference coefficients (delta MFCCs), its acceleration (delta delta MFCCs), delta power, and delta delta power, and they were calculated every 10 msec. The distribution of the acoustic features was modeled using 32 mixtures of diagonal covariance Gussians for the

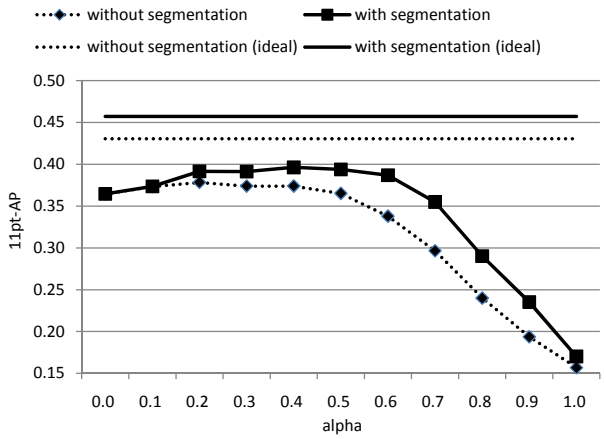

Fig. 3. *Retrieval performance with dynamic segmentation of spoken document for all 39 queries.*

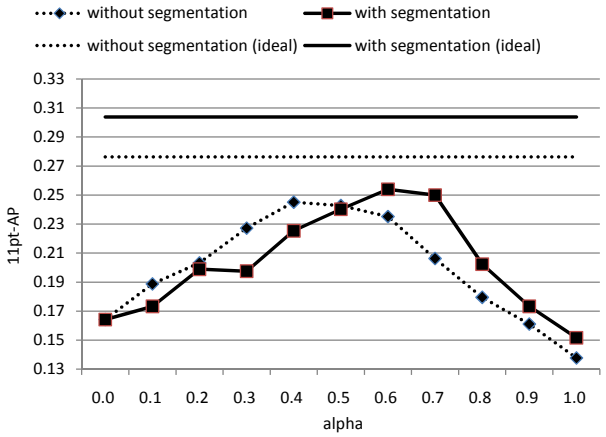

Fig. 4. *Retrieval performance with dynamic segmentation of spoken document for 11 queries including OOV words.*

HMMs. In addition, we prepared the word trigram based LM with 17k kinds of words in the recognition dictionary.

Table I shows the recognition performance on the CSJ lectures. "Corr." and "Acc." mean the word correct rate and accuracy rate, respectively. "OOV" is the OOV rates for only the all query sentences. Note that this is not OOV rate for the spoken lectures. "# of OOV" shows the number of queries including any OOV words.

C. Evaluating Metric

For our evaluation metric, we used "11pt-AP," which is obtained by averaging the following interpolated 11-points average precision (denote as "AP") over the 39 queries[10].

$$IP(x) = \max_{x \geq R_i} P_i, AP = \frac{1}{11} \sum_{i=0}^{10} IP(\frac{i}{10}) \qquad (3)$$

where R_i and P_i are the recall and precision, respectively, up to the i-th retrieved documents. In practice, we retrieved 1000 documents for each query to calculate the AP.

D. Experimental Results

Figure 3 and Figure 4 show that 11pt-AP values changing the value of α on the two cases: with automatic document segmentation and without it. In addition, Figure 3 and Figure 4 represents the result for all the 39 queries and only the 11 queries including OOV words, respectively.

The value of 11pt-AP with $\alpha = 0.0$ and the lines tagged by "without segmentation" are regard as "baseline" in this paper. $\alpha = 0.0$ means that the search engine does not use any WI. α is fixed to all queries excepting the straight line with "Ideal" tag. They mean the 11pt-AP values when the suitable α for each query is set.

The best 11pt-AP value is 0.401 ($penalty = 0.5, \alpha = 0.4$) for all the queries when the document segmentation is used and theα is fixed to all queries. The value is higher than the one of baseline (0.378).

In addition to this, the OOV query set gets the best 11pt-AP values at higher α comparing with the 11pt-AP value of all the query set. Furthermore, the α value obtaining the maximum 11pt-AP is 0.6 when the document segmentation is used for only the OOV queries. Its value is higher than the α value of the case without segmentation. This claims that the WEB collection with the document segmenatation can get more suitable WEB pages than the collection without the segmentation.

IV. Conclusions

In this paper, we have shown the effectiveness of document expansion using WEB pages which were collected by the document segmenataion for the spoken lecture retrieval task. In an experiment, our technique achieved an relative improvement of 10.2% on the 11pt-AP based on interpolated 11-points average precision for the all queries on the test collection and 61.0% for the 11 queries that include OOV words in comparison of the 11pt-AP without the document expansion. These results demonstrate that document expansion using WEB pages is useful for spoken document retrieval.

In future work, we are going to develop novel WEB collection techniques, including how to compose WEB search queries, how to combine the retrieval score of RI and WI, and so on

References

[1] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.
[2] C. Ng et al., "Experiments in spoken document retrieval using phoneme n-grams," *Speech Communication*, pp. 61–77, 2000.
[3] Y. Itoh et al., "An integration method of retrieval results using plural subword models for vcabulary-free spoken document retrieval," in *Proc. of INTERSPEECH2007*, 2007, pp. 2389–2392.
[4] Y.C. Pan et al., "Subword-based Position Specific Posterior Lattices (S-PSPL) for Indexing Speech Information," in *Proc. of the INTERSPEECH 2007*, 2007, pp. 318–321.
[5] A. Singhal and F. Pereira, "Document Expansion for Speech Retrieval," in *Proc. of ACM SIGIR'99*, 1999, pp. 34–41.
[6] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
[7] M. Utiyama and H. Isahara, "A Statistical Model for Domain-Independent Text Segmentation," in *Proc.of the 9th ECACL*, 2001, pp. 491–498.
[8] A. Takano et al., "Generic Engine for Transposable Association (GETA)," http://nii.ac.jp/geta/english.html, (refered on) .
[9] A. Singhal et al., "Pivoted document length normalization," in *Proc. of ACM SIGIR' 96*, 1996, pp. 21–29.
[10] T. Akiba et al., "Test Collections for Spoken Document Retrieval from Lecture Audio Data," in *Proceeding of the 6th edition of the Language Resources and Evaluation Conference (LREC)*, 2008.