Selection of Best Match Keyword using Spoken Term Detection for Spoken Document Indexing

Kentaro Domoto*, Takehito Utsuro*, Naoki Sawada† and Hiromitsu Nishizaki†
* Graduate School of Systems and Information Engineering,
University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-0006 Japan
† Department of Education, Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Kofu-shi, Yamanashi 400-8511 Japan
E-mail: {sawada,nisizaki}@alps-lab.org Tel: +81-55-220-8361

***Abstract*—This paper presents a novel keyword selection-based spoken document-indexing framework that selects the best match keyword from query candidates using spoken term detection (STD) for spoken document retrieval. Our method comprises creating a keyword set including keywords that are likely to be in a spoken document. Next, an STD is conducted for all the keywords as query terms for STD; then, the detection result, a set of each keyword and its detection intervals in the spoken document, is obtained. For the keywords that have competitive intervals, we rank them based on the matching cost of STD and select the best one with the longest duration among competitive detections. This is the final output of STD process and serves as an index word for the spoken document. The proposed framework was evaluated on lecture speeches as spoken documents in an STD task. The results show that our framework was quite effective for preventing false detection errors and in annotating keyword indices to spoken documents.**

I. Introduction

In recent years, information technology environment have evolved such that numerous audio and multimedia archives, such as video archives and digital libraries, can be easily accessed. In particular, a rapidly increasing number of spoken documents, such as broadcast programs, spoken lectures, and recordings of meetings, are archived; some of the archived documents are accessible on the Internet. Although the need to retrieve such spoken information is growing, at present there is no effective retrieval technique available; thus, the development of technology for retrieving such information has become increasingly important.

In the Text REtrieval Conference (TREC) Spoken Document Retrieval (SDR) track hosted by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) in the second half of the 1990s, many SDR studies that used English and Mandarin broadcast news documents were presented [1]. TREC SDR is an ad-hoc retrieval task that retrieves spoken documents that are highly relevant to a user query.

On the other hand, the Spoken Document Processing Working Group, which is part of the special interest group of spoken language processing (SIG-SLP) of the Information Processing Society of Japan, has already developed prototypes of SDR test collections: the Corpus of Spontaneous Japanese (CSJ) Spoken Term Detection test collection [2] and CSJ Spoken Document Retrieval test collection [3]. The target documents of both test collections are spoken lectures in CSJ [4]. In addition, SDR and spoken term detection (STD) tasks were proposed in NTCIR-9 [5] and NTCIR-10 [6] conferences, and many research groups joined the task and presented their frameworks on SDR and STD.

If spoken documents related to a query are specified by an SDR technique, it is very difficult to find the highly relevant speech sections of the retrieved spoken documents without listening to all the speeches. Furthermore, an SDR technique may highly rank spoken documents in which keywords constituting the query are never uttered. This may make an SDR user edgy. Therefore, studies on STD, which can indicate speech intervals where the query term is uttered in spoken documents, became popular after NIST initiated the STD project with a pilot evaluation and workshop [7] in 2006. If target keywords are specified in a retrieved spoken document by an STD technique, an SDR user can easily create a cueing of the retrieved speech using the specified term and can listen to the specific speech interval.

Therefore, a combination of SDR and STD technologies is very useful in speech information access. In general, SDR and STD methods use automatic speech recognition (ASR) technology for transcribing a target speech before a query search process. Here, one of the difficulties in SDR and STD in the search occurs due to erroneous transcriptions. Another difficulty is also in the search for terms in a vocabulary-free framework, as the search terms are not known prior to the ASR system being used. Most SDR and STD studies focus on the out-of-vocabulary (OOV) and the ASR error problems [8], [9]. For example, STD techniques that use entities such as subword lattice and confusion network (CN) have been proposed [10], [11].

This paper proposes a novel framework to index a spoken document using the STD technique for SDR. In a similar study, Takigami et al. [12] proposed an SDR method using an STD framework. Takigami's method first conducts STD for each keyword appearing in the given query topic; then, all the detections are used to calculate the relevance of the retrieved document to the topic. In contrast, our proposed method filters the detections for each keyword.

In a previous study [13], Natori et al. reported on STD from spontaneous spoken lectures using a phoneme transition

978-616-361-823-8

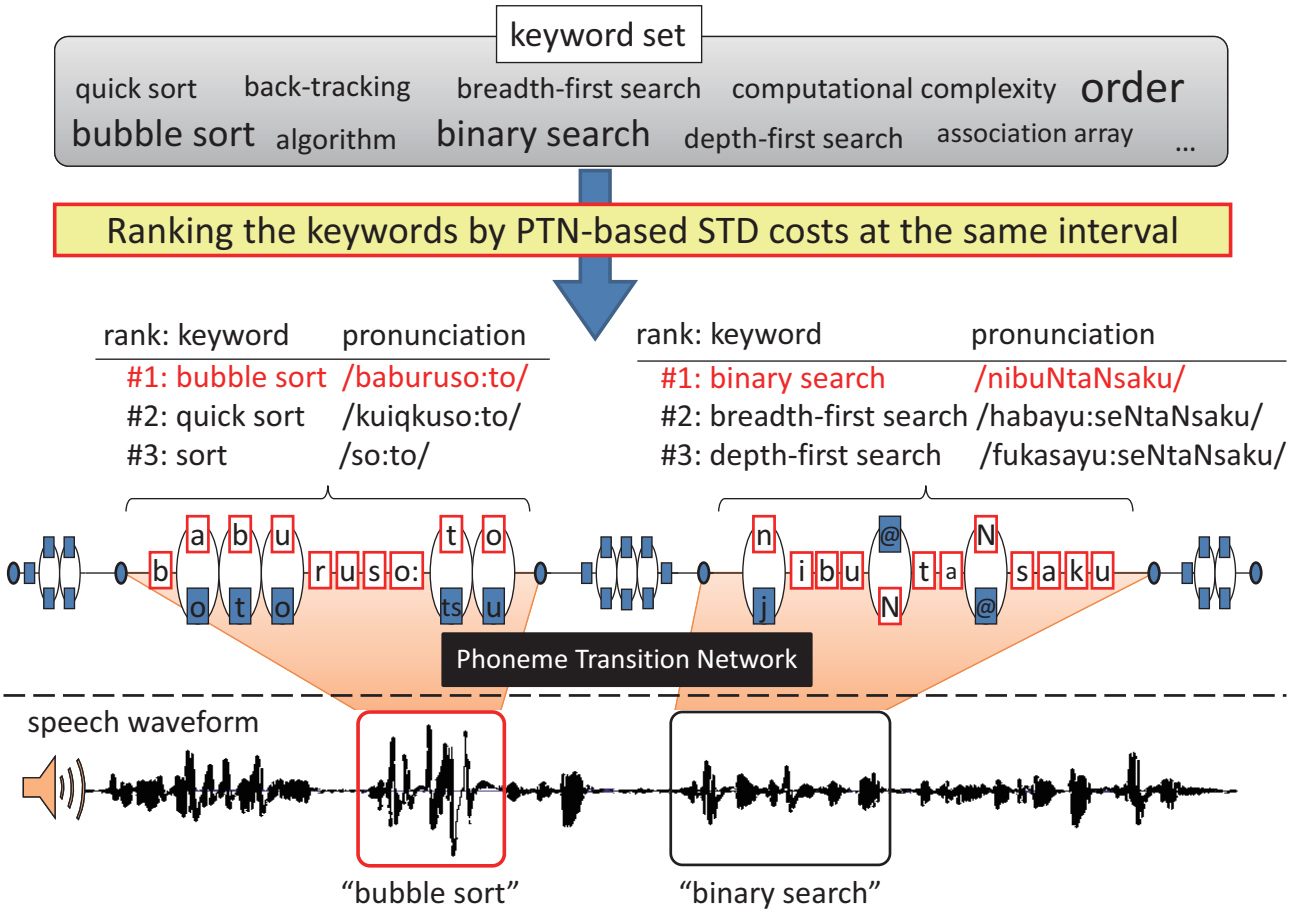

Fig. 1. An index framework of intervals in a spoken document using spoken term detection.

network (PTN)-formed index derived from multiple ASR systems' 1-best hypotheses. PTN-based indexing is based on an idea of CN generated from an ASR system. CN-based indexing for STD is known as a powerful indexing method. The PTN-formed index is generated by merging the phoneme sequences of ASR systems' outputs to a single CN. In this study, we use this STD engine to index a spoken document. Although Natori's STD method was robust for misdetections, it raised the number of false detections because it has a more complicated CN structure created by a single ASR system. In addition, the STD method used a subword-based (phoneme-based) matching between a query and an index. Therefore, some query terms that have the same or similar pronunciations are essentially detected at the same speech intervals in the spoken document. This is a weak point of general STD techniques that have already been proposed. This means that more than one word is indexed at the same position. This over-detection problem is a fatal issue on the STD-based indexing for SDR.

Figure 1 shows the matching examples of two indexing words. In the example, a keyword “bubble sort” is matched to a speech interval, but the other keywords, “quick sort” and “sort,” are also matched to the same interval or a part of the interval. Therefore, it is necessary to avoid these overlapped detections. First, our indexing framework creates a keyword set including keywords that are likely to appear in a spoken document. Next, the STD engine detects the speech interval candidates of all the keywords. Then, a set of each keyword and its detection intervals in the spoken document is obtained. For the keywords that have competitive intervals, we rank the ones based on the matching cost of STD. We select the best one with the longest duration among all the competitive detections. This is the final output of the STD process, and it serves as an index word of the spoken document. In this paper, we compare three kinds of candidate selection methods, including the baseline and the two proposed methods.

We evaluate the proposed framework on lecture speeches as spoken documents on an STD task, and the results show that the proposed framework is quite effective in annotating keyword indices to spoken documents.

II. Spoken Term Detection Engine and ASR

We employ the STD engine [13] that uses subword-based CN. We use a PTN-formed index derived from multiple ASR systems' 1-best hypotheses and an edit distance-based dynamic time warping (DTW) framework to detect a query term. This study employs 10 types of ASR systems; the same decoder was used for all types. Two types of acoustic

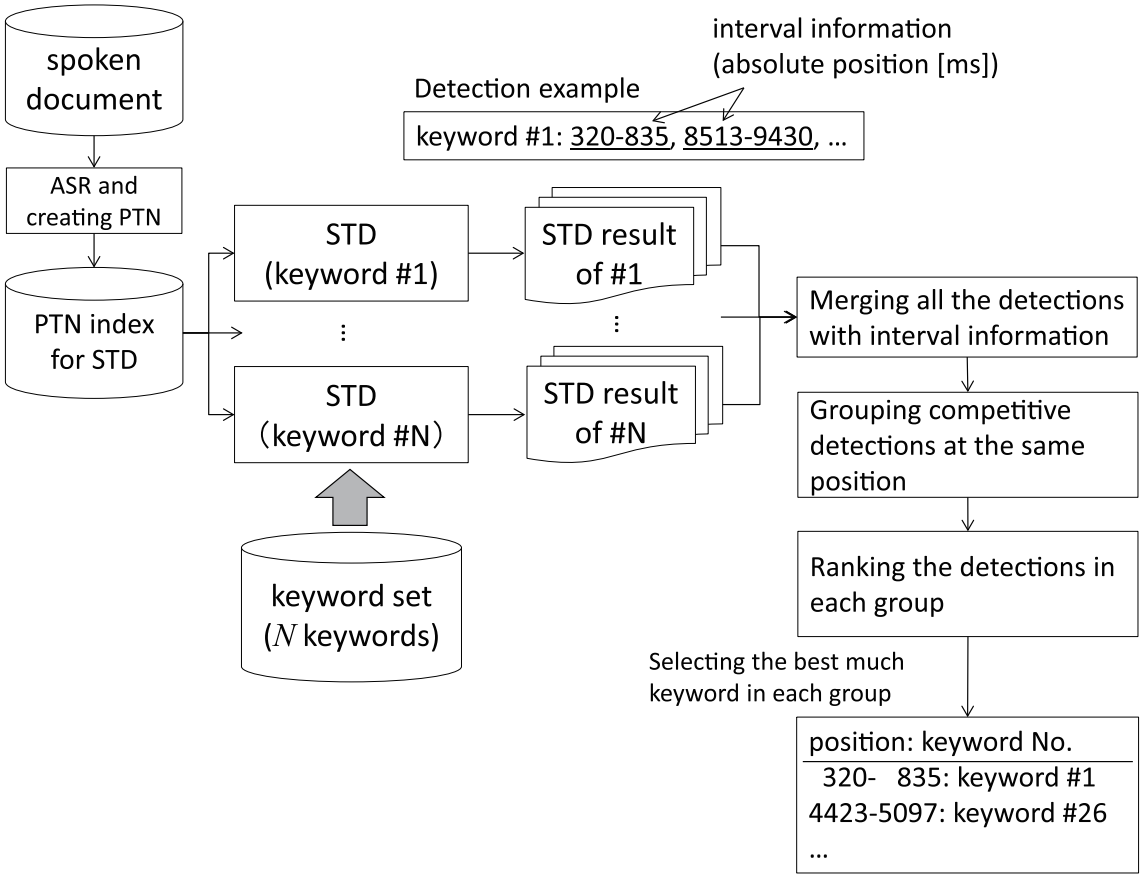

Fig. 2. Workflow on indexing a spoken document using the proposed method.

models and five types of language models were prepared. The multiple ASR systems can generate the PTN-formed index by combining subword (phoneme) sequences from the output of these ASR systems into a single CN. The details of the STD engine are explained in [13]. The STD engine includes some parameters for DTW. This study uses the STD engine with the false detection parameters of "Voting" and "AcwWith," which received the best STD performance on the evaluation sets [13].

Julius ver. 4.1.3 [14], an open source decoder for ASR, is used in all systems. Acoustic models are triphone-based (Tri.) and syllable-based (Syl.) Hidden Markov Models (HMMs), both of which are trained on spoken lectures in CSJ [4]. Language models are word-based as well as character-based trigrams as follows:

- WBC : word-based trigram where words are represented by a mix of Chinese characters, and Japanese Hiragana and Katakana.
- WBH : word-based trigram where all words are represented only by Japanese Hiragana. Words comprising Chinese characters and Japanese Katakana are converted into Hiragana sequences.
- CB : character-based trigram where all characters are represented by Japanese Hiragana.
- BM : character-sequence-based trigram where the unit of language modeling comprises two Japanese Hiragana characters.
- Non : no language model is used. Speech recognition without any language model is equivalent to phoneme (or syllable) recognition.

Each model was trained using CSJ transcriptions.

The training conditions of all acoustic and language models and ASR dictionary are the same as in STD/SDR test collections used in the NTCIR-9 Workshop [5].

To evaluate our framework, we use 11 high-quality lecture speeches in CSJ and a low-quality (very noisy) simulated classroom lecture speech, recorded at the University of Yamanashi. The word-correct and accuracy rates for the 11 lecture speeches are about 79% and 75%, respectively, when the ASR system with the combination of WBC and Tri. models is used to transcribe them. On the other hand, for the classroom speech, the word-correct and accuracy rates are 26% and 9%, respectively [14].

III. Selection of the Best Match Keyword

Figure 2 shows an outline of the keyword indexing of a spoken document by selecting the best match keyword from competitive detections using the STD technique. The details are explained in the following sections.

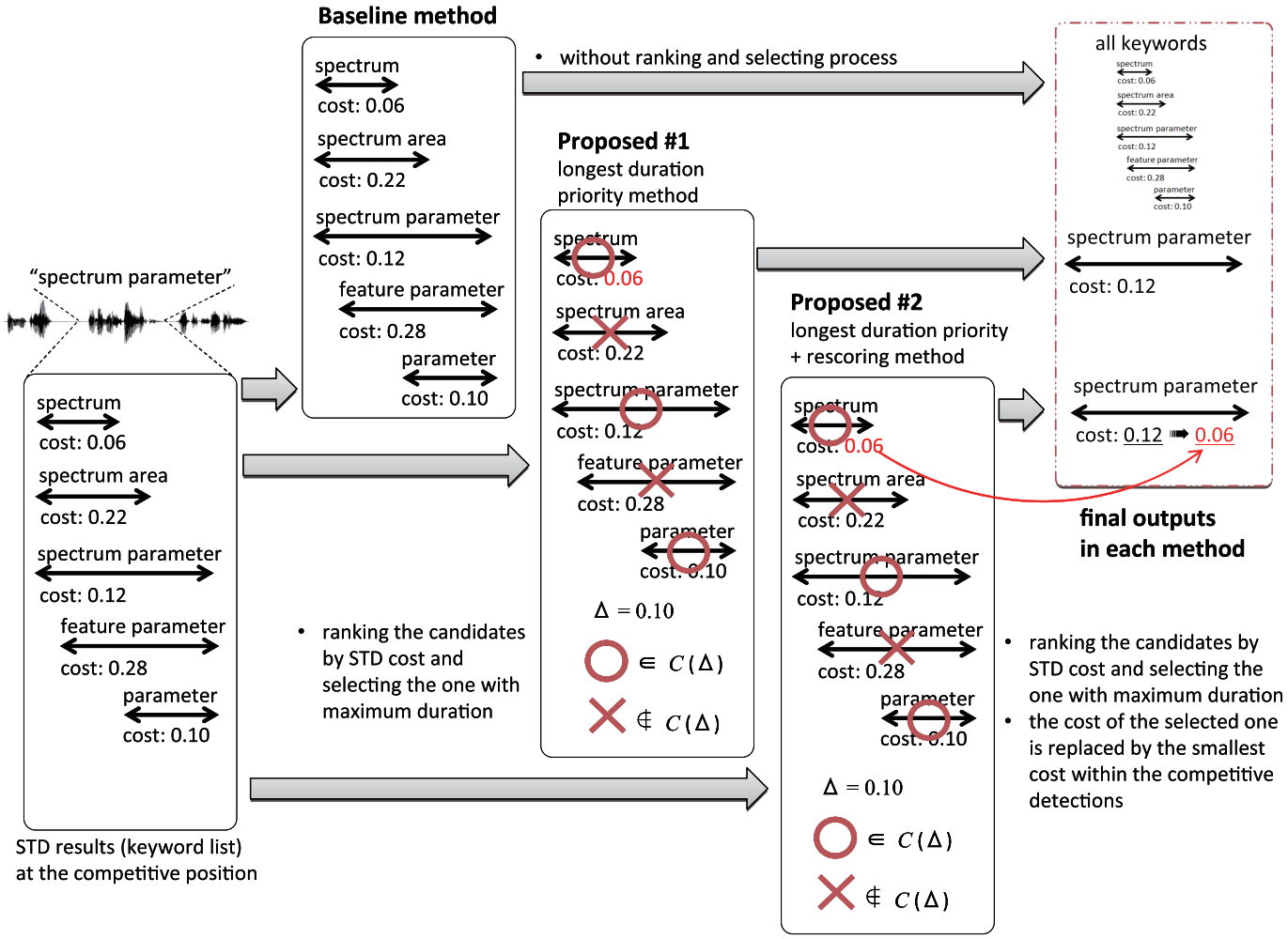

Fig. 3. Three selection methods from a competitive interval.

A. Keyword Set

In this study, we manually create a keyword set for each spoken document. First, we use a technical term extraction tool "TermExtract" [1] [15] to obtain keyword (technical term) candidates from the transcriptions of the spoken documents. Out of the extracted candidates, we manually select the keywords that are used as queries for STD.

B. Grouping Competitive Detections

A spoken document is automatically transcribed by the 10 types of ASR systems and the PTN-formed index is created from the transcriptions. STD engine outputs intervals where a query term is likely to be uttered. First, we conduct STD for all keywords as query terms. The detection result set is obtained for each keyword. The detection intervals of all keywords are merged. Some keywords are detected at the same position as other keywords, and some keywords are detected as part of the intervals of other keywords; we call these "competitive detections."

Next, we transitively group the overlapped detections, which are at the same position or whose detected intervals are partially overlapped. We define the grouped detections as a competitive set C, where the set C is constructed transitively by collecting a detection which overlaps at least one member detection within C other than itself.

C. Keyword Selection from Competitors

Figure 3 shows the three indexing methods for a spoken document. "Baseline method" is the same as a typical STD scheme that selects all the keywords belonging to C. In other words, the baseline outputs and indexes all keywords that have STD costs below a previously set threshold. "Proposed method #1," the longest-duration priority method, selects the keyword that has the longest duration among all the keywords in the competing group and that also has STD costs within the cost range dynamically set by the smallest cost in the group. "Proposed method #2," the longest-duration priority and rescoring method, is similar to the proposed method #1. The only difference between the two proposed methods is the STD cost attached to the selected keyword.

[1] http://gensen.dl.itc.u-tokyo.ac.jp/ This page is written in Japanese only.

We explain the details of the two proposed methods in the following sections.

1) Longest-duration priority: We define a quadruplet, which comprises a keyword w, the start time of its detection interval t, the end time of the one t', and the STD matching cost of the keyword $cost$. A competitive detection set C comprising n quadruplets $\langle w, t, t', cost \rangle$ is defined as follows:

$$C = \{\langle\ w_1, t_1, t'_1, cost_1\rangle, \cdots, \langle w_n, t_n, t'_n, cost_n\rangle\ \}$$

The C in the case of Figure 3 is represented as follows:

$$\begin{aligned} C \quad = \{ \quad & \langle \text{spectrum}, t_1, t_3, 0.06\rangle, \\ & \langle \text{spectrum area}, t_1, t_4, 0.22\rangle, \\ & \langle \text{spectrum parameter}, t_1, t_5, 0.12\rangle, \\ & \langle \text{feature parameter}, t_2, t_5, 0.28\rangle, \\ & \langle \text{parameter}, t_3, t_5, 0.10\rangle\ \} \end{aligned}$$

where $t_1 < t_2 < t_3 < t_4 < t_5$. In this method, we first find the quadruplet that has the smallest matching cost from C:

$$\langle w_{min}, t, t', cost_{min}\rangle.$$

In the case of Figure 3, the following quadruplet:

$$\langle \text{spectrum}, t_1, t_3, 0.06\rangle.$$

is selected.

Next, a candidate set $C(\Delta)$ for indexing is created by filtering quadruplets in C based on the cost-range Δ. The quadruplets in $C(\Delta)$ have the STD cost less than $(cost_{min} + \Delta)$ as follows:

$$C(\Delta) = \{\langle w, t, t', cost\rangle\ \in\ C \mid cost \leq (cost_{min} + \Delta)\}$$

For example, in Figure 3, if $\Delta = 0.10$, $C(\Delta = 0.10)$ is represented as follows:

$$\begin{aligned} C(\Delta = 0.10) \quad = \{ \quad & \langle \text{spectrum}, t_1, t_3, 0.06\rangle, \\ & \langle \text{spectrum parameter}, t_1, t_5, 0.12\rangle, \\ & \langle \text{parameter}, t_3, t_5, 0.10\rangle\ \} \end{aligned}$$

Finally, we select the quadruple $\langle w_{ld}, t, t', cost_{ld}\rangle$, which has the longest duration (ld) from $C(\Delta)$ when a duration of a detected keyword is defined as $t' - t$. The keyword w_{ld} is outputted as the STD result, and used as an indexing word for a spoken document. In the example of Figure 3, the following quadruple is the final output:

$$\langle \text{spectrum parameter}, t_1, t_5, 0.12\rangle$$

Competitive detections, whose intervals overlap with the selected detection, are removed from C. Then, the above process is repeated until C becomes an empty set.

2) Longest-duration priority and rescoring: In general, the more phonemes a keyword is decomposed into, the greater the STD matching cost outputted by an STD engine is. Longer keywords outputted by the longest-duration priority method have higher STD cost in comparison with shorter keywords. Considering this situation, we improve the longest-duration priority method by simply replacing the STD matching cost of the longest-duration priority method with the smallest cost in the competitive group $C(\Delta)$. The quadruple output using this method is represented as follows where $cost_{min}$ denotes the smallest cost in the competitive group $C(\Delta)$:

$$\langle w_{ld}, t, t', cost_{min}\rangle.$$

For example, in the case of the example in Figure 3, the quadruple is shown as follows:

$$\langle \text{spectrum parameter}, t_1, t_5, 0.06\rangle$$

IV. Evaluation

A. Experimental Setup

We evaluated the three selection methods: the baseline (same as STD), the longest-duration priority method (proposed #1), and the rescoring method (proposed #2) on an STD task, which normally evaluates keyword-detection performance of a query set on a spoken document collection. Although we eventually aim to improve the indexing accuracy of a spoken document, the spoken document indexing based on our proposed techniques is affected by STD performance. In this paper, therefore, the evaluation task was to measure STD performance for a spoken document as an indexing assessment because we generated keyword sets for each spoken document. An STD query for a spoken document was all the keywords in the keyword set, which is created by the process described in Section III-A, for the spoken document.

The evaluation speech data comprised 11 lectures (five academic lecture speeches, denoted as "CSJ-AL," and six simulated lecture speeches, denoted as "CSJ-SL") in CSJ and the one simulated classroom lecture speech recorded at the University of Yamanashi (denoted as "UY-CL") [14]. TABLE I shows the average number of keywords (query terms) and keyword occurrences per lecture. The numbers enclosed in parentheses depict the number of OOV keywords or keyword occurrences. OOV keywords are not registered in the ASR (WBC/Tri. system) dictionary.

he ASR performance (word-correct and accuracy rates) of the evaluation data is also represented in TABLE I. Although CSJ lecture speeches were of high acoustic quality with low noise, the ASR performance of CSJ-SL was lower than that of CSJ-AL because of topical adaptivity to the language model used in the ASR system. It was very difficult to speech-recognize UY-CL because of its low-quality (very noisy) and topical adaptivity to the language model.

The STD cost range Δ was set to 0.10 in this experiment.

TABLE I
EXPERIMENTAL CONDITIONS IN THE EVALUATION DATA. NUMBERS ENCLOSED IN PARENTHESES SHOW THE NUMBER OF OOV KEYWORDS OR KEYWORD OCCURRENCES.

		CSJ 11 lectures	CSJ-AL	CSJ-SL	UY-CL
Number of keywords/lecture		68 (7)	83 (3)	55 (11)	96 (8)
Number of keyword occurrences/lecture		253 (15)	408 (5)	124 (24)	587 (22)
ASR rate [%]	Correct	79.2	82.9	74.5	26
	Accuracy	74.6	79.4	69.7	9

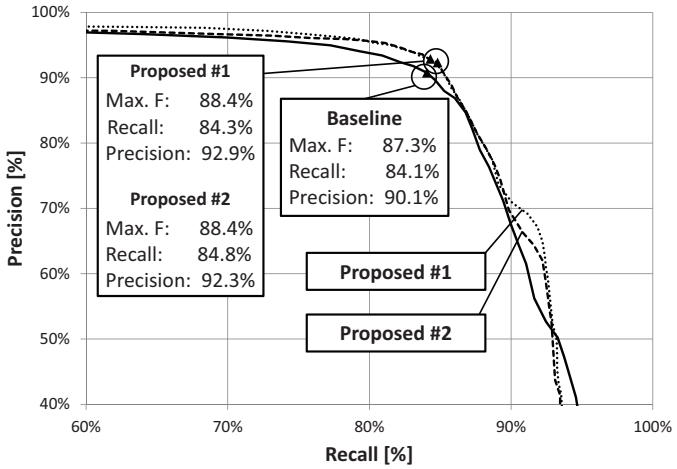

Fig. 4. Recall-precision curves for all the keywords evaluated on all CSJ 11 lectures.

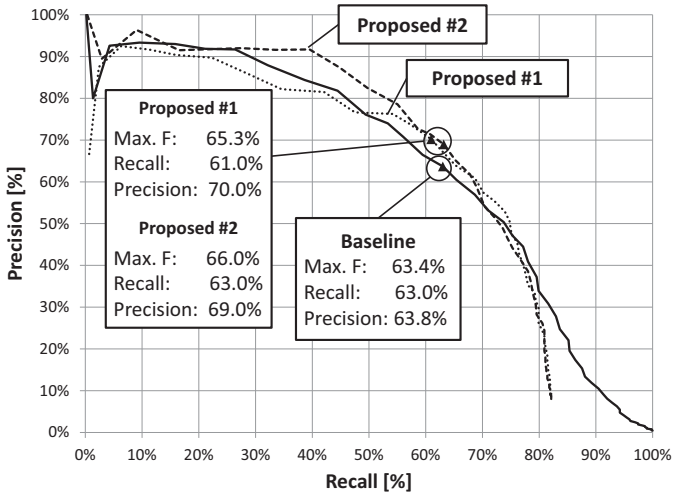

Fig. 5. Recall-precision curves for all the keywords evaluated on UY-CL.

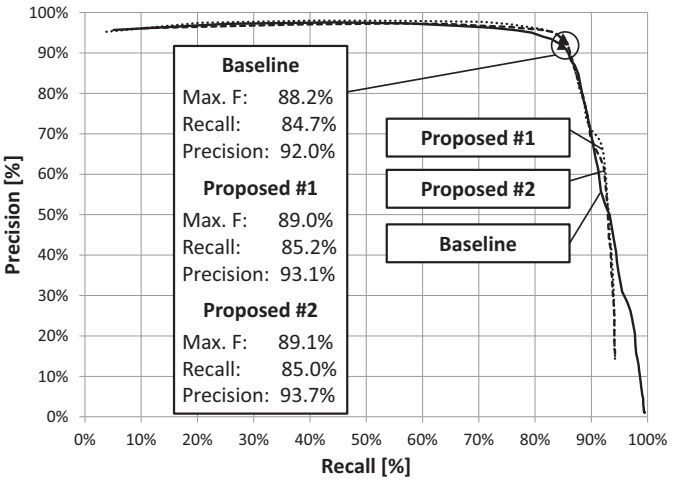

Fig. 6. Recall-precision curves for IV keywords evaluated on all 11 CSJ lectures.

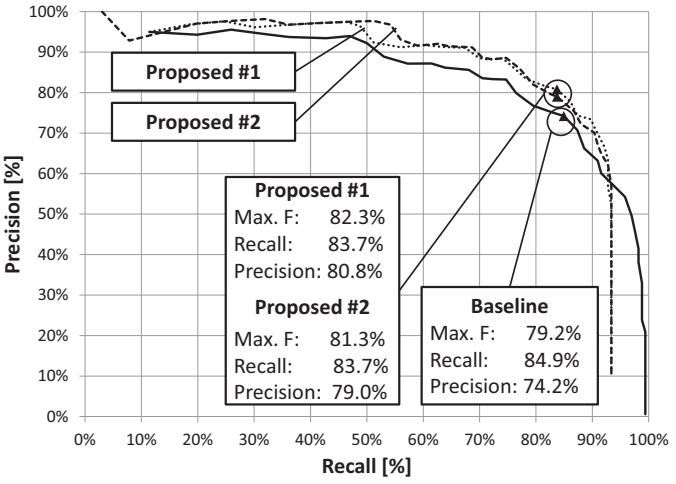

Fig. 7. Recall-precision curves for OOV keywords evaluated on all 11 CSJ lectures.

B. Evaluation Metrics

The evaluation metrics used in this study are recall, precision, and F-measure. These measurements are frequently used to evaluate the information retrieval performance, and they are defined as follows:

$$\text{Recall} = \frac{N_{corr}}{N_{true}}$$

$$\text{Precision} = \frac{N_{corr}}{N_{corr} + N_{spurious}}$$

$$\text{F-measure} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

Here N_{corr} and $N_{spurious}$ are the total number of correct and spurious (false) keyword detections, and N_{true} is the total number of true keyword occurrences in the speech data. F-measure values for the optimal balance of $Recall$ and $Precision$ values are denoted by "Max. F" in the evaluation graphs.

The STD performance for the keyword sets can be displayed by a recall-precision curve, which is plotted by changing the

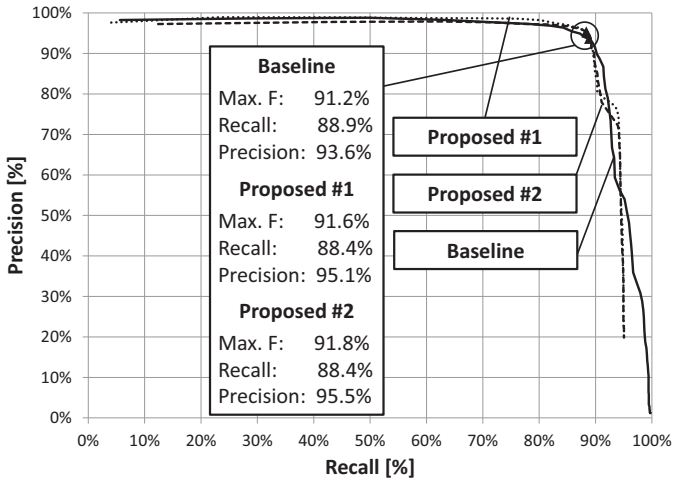

Fig. 8. Recall-precision curves for IV keywords evaluated on CSJ-AL.

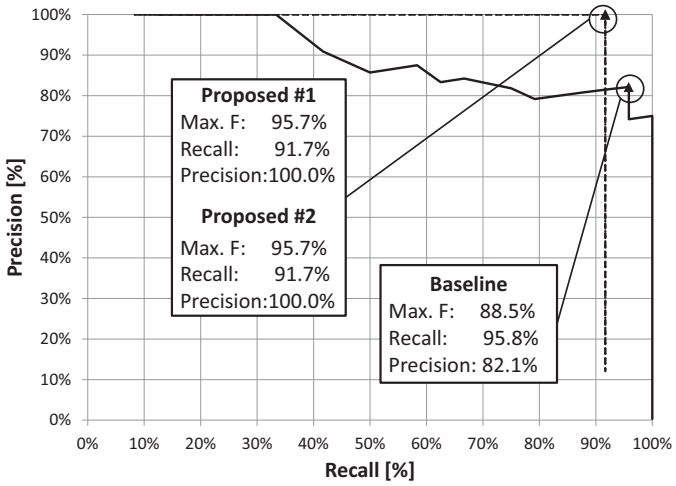

Fig. 9. Recall-precision curves for OOV keywords evaluated on CSJ-AL.

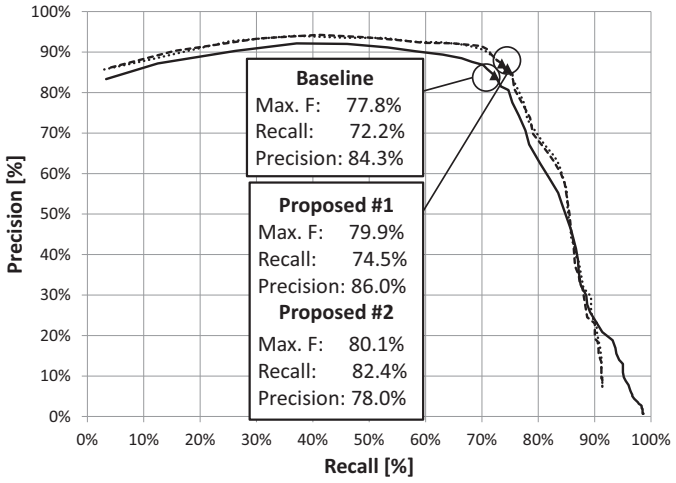

Fig. 10. Recall-precision curves for IV keywords evaluated on CSJ-SL.

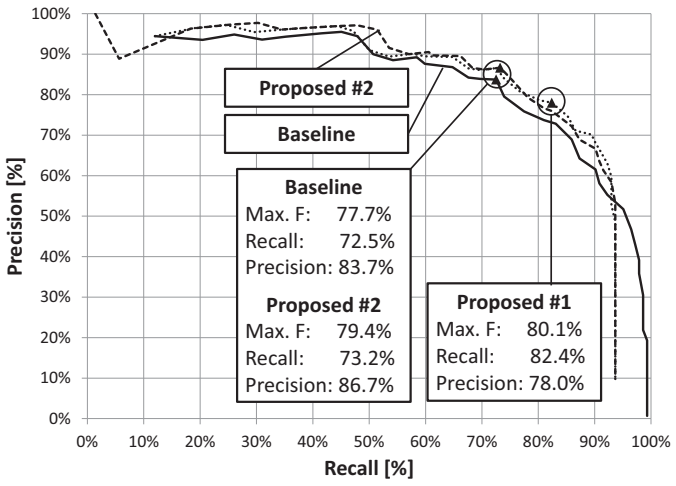

Fig. 11. Recall-precision curves for OOV keywords evaluated on CSJ-SL.

threshold θ value on the STD costs of outputted keywords by the proposed methods.

C. Experimental results and discussion

Figures 4 and 5 show the recall-precision curves for all 11 lectures in CSJ and UY-CL, respectively. Figures 6 through 11 also show recall-precision curves for the 11 CSJ lectures, CSJ-AL and CSJ-SL, for each kind of keyword (in-vocabulary (IV) or OOV keywords). As shown in these results, the precision values of the proposed methods are lower than the baseline in the high-recall region (over 90% of recall in the CSJ lectures and over 70% of recall in UY-CL), because the proposed methods output only one keyword in a competitive interval. However, our methods outperformed the baseline in the other regions. Proposed methods #1 and #2 improved the baseline STD performance for all evaluation speeches.

First, we discuss the relation between the ASR performance and the effectiveness of the proposed methods. As shown in Figures 4 – 11, all the curves except for IV keywords in CSJ-AL (Figure 8) are better than the baseline curves in maximum F-measures. In particular, the proposed methods produced greater improvements for CSJ-AL with OOV keywords, CSJ-SL with OOV keywords, and UY-CL compared with CSJ-AL and CSJ-SL with IV keywords. OOV keywords cannot be correctly transcribed as words by an ASR system. In addition, the conversion of an OOV keyword utterance into a correct subword (phoneme) sequence is difficult. Therefore, the quality of the phonetic transcription of OOV keywords was worse than that of IV keywords, and STD performances were also worse for OOV keywords than for IV keywords. Furthermore, UY-CL was very difficult to speech-recognize, as shown in TABLE I. Not all keywords for UY-CL were correctly transcribed to the phonetic transcriptions; hence, the STD performance for UY-CL was lower than that for the CSJ lectures. However, our proposed methods were effective for the lower-quality transcriptions of the lectures. An STD for speeches that are difficult to speech-recognize is likely to generate false detection errors because the PTN-formed index for the STD includes an unsure phoneme sequence, which is falsely matched to incorrect query terms. Our methods could

control the false detections by selecting the best keyword from the candidates in the same competitive intervals. On the other hand, IV keywords in CSJ-AL that have good ASR performances are easily transcribed to the confident phoneme sequences. Because the baseline STD could exert the efficient performance, the proposed methods could not outperform the baseline in CSJ-AL with IV keywords.

Next, comparing the proposed methods #1 and #2, no difference exists in any of the test speeches on the maximum F-measure metric except for UY-CL. With UY-CL, which was difficult for an ASR system to transcribe, proposed method #2 obtained the best STD performance, as shown in Figure 5. Although an STD match cost is likely to be higher when an ASR performance is worse, the best selected keyword, which has the most phonemes among the competitive candidates, may have good reliability. Proposed method #2 ensures this by providing the lowest cost to the selected keyword. In recall-precision curves for UY-CL, although the false detections start to increase from the recall rate of approximately 40%, method #2 controlled it. The same is equally true of the curves for CSJ-SL with OOV keywords in Figure 11. This asserts that proposed method #2 is effective when a speech that is difficult to speech-recognize and includes OOV keywords is indexed.

V. Conclusion

This paper proposed a novel keyword selection method using the STD framework for spoken document indexing. In our framework, first we create a keyword set, comprising keywords that are likely to appear in a target spoken document, and then, the STD process was conducted for all keywords on the spoken document. Next, all detections were classified into competitor groups based on the speech interval information of the detected keywords. Keyword candidates that had competitive intervals in a group were ranked on the basis of STD cost. Finally, the best keyword was selected based on the duration and its STD cost was rescored. We evaluated the proposed selection methods on the STD task. The experimental results showed that the proposed methods were quite effective and also robust for indexing a spoken document whose transcription has many ASR errors and OOV keywords.

In this paper, we manually created a keyword set for STD query. In future studies, we are going to develop a method for automatically creating the keyword set using Web pages related to a target spoken document.

VI. Acknowledgements

This study was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 26282049.

References

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," in *Proceedings of the Text REtrieval Conference (TREC) 8*, 2000, pp. 16–19.

[2] Y. Itoh, H. Nishizaki, X. Hu, H. Nanjo, T. Akiba, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita, and K. Aikawa, "Constructing Japanese Test Collections for Spoken Term Detection," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH2010)*. ISCA, 2010, pp. 677–680.

[3] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamanashita, and K. Itou, "Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data," *Journal of Information Processing*, vol. 17, pp. 82–94, 2 2009.

[4] K. Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[5] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for Spoken Documents Task in NTCIR-9 workshop," in *Proceedings of the 9th NTCIR Workshop Meeting*, 2011, pp. 223–235.

[6] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamanashita, "Overview of the NTCIR-10 SpokenDoc-2 Task," in *Proceedings of the 10th NTCIR Conference*, 2013, pp. 573–587.

[7] "The spoken term detection (STD) 2006 evaluation plan," 2006, http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.%pdf.

[8] K. Iwata, K. Shinoda, and S. Furui, "Robust Spoken Term Detection Using Combination of Phone-based and Word-based Recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*. ISCA, 2008, pp. 2195–2198.

[9] B. Logan and J.-M. V. Thong, "Confusion-based query expansion for OOV words in spoken document retrieval," in *Proceedings of ICSLP 2002*, 2002, pp. 1997–2000.

[10] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH2007)*. ISCA, 2007, pp. 2393–2396.

[11] S. Meng, J. Shao, R. P. Yu, J. Liu, and F. Seide, "Addressing the out-of-vocabulary problem for large-scale chinese spoken term detection," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*. ISCA, 2008, pp. 2146–2149.

[12] T. Takigami and T. Akiba, "Open Vocabulary Spoken Content Retrieval by front-ending with Spoken Term Detection," in *Proceedings of the 5th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013)*, 2013, pp. 1–6.

[13] S. Natori, Y. Furuya, H. Nishizaki, and Y. Sekiguchi, "Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers' Outputs," *Journal of Information Processing*, vol. 21, no. 2, pp. 176–185, 2013.

[14] C. Yonekura, Y. Furuya, S. Natori, H. Nishizak, and Y. Sekiguchi, "Evaluation of the Usefulness of Spoken Term Detection in an Electronic Note-Taking Support System," in *Proceedings of the 5th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013)*, 2013, pp. 1–4.

[15] H. Nakagawa and T. Mori, "Automaic Term Recognition based on Statistics of Compound Nouns and their Components," *Terminology*, vol. 9, no. 2, pp. 201–209, 2003.