Spoken Term Detection Using Spoken Document Index Based on Keywords Collected from Automatic Speech Recognition Result

Kentaro Domoto and Takehito Utsuro
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
Email: s1420804_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Naoki Sawada and Hiromitsu Nishizaki
Department of Education, Interdisplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu, Japan
Email: {sawada, nisizaki}_@_alps-lab.org

***Abstract*—This paper presents a novel spoken document indexing framework for Spoken Term Detection (STD). Our proposed method utilizes an STD method for making an index from keywords collected from outputs from automatic speech recognition systems. The STD method is conducted for all the keywords as query terms; 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 real 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.**

***Index Terms*—keyword collection, spoken document indexing, spoken term detection**

I. Introduction

In recent years, information technology environments 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], NTCIR-10 [6] and NTCIR-11 [7] 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 [8] 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. Therefore, the difficulty in SDR and STD in the search occurs due to erroneous transcriptions and the search for terms in a

Manuscript received May 3, 2015; revised September 3, 2015.

doi: 10.18178/ijsps.4.4.282-288

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 [9], [10]. For example, STD techniques that use entities such as subword lattice and Confusion Network (CN) have been proposed [11], [12].

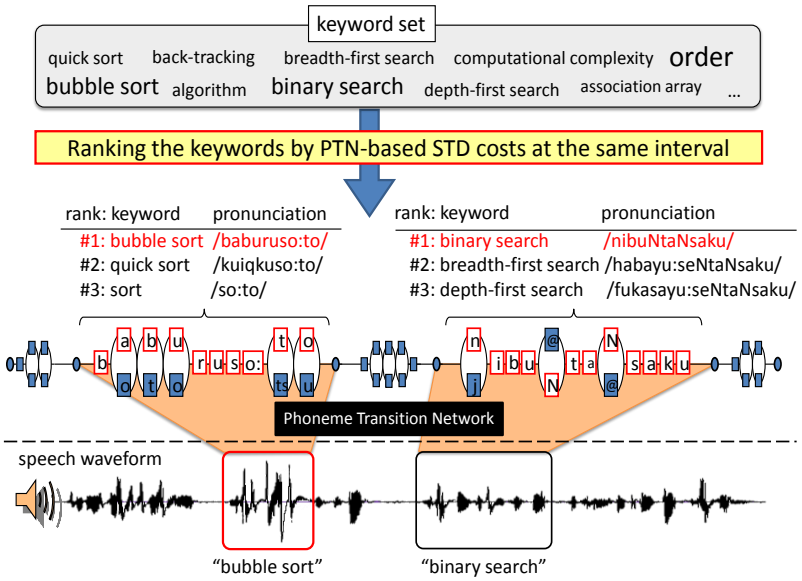

Figure 1. An index framework of intervals in a spoken document using a spoken term detection method.

In a previous study [13], Natori *et al.* reported on STD from spontaneous spoken lectures using a phoneme transition 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 miss-detections, 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 STD and SDR.

This paper proposes a novel framework to index a spoken document using an STD technique for STD and SDR. We call it "Selection of Best Matched Keyword (SBMK)" indexing method. The SBMK method uses a keyword list that is made from an ASR system's output. Our method previously attaches a keyword in a list on suitable intervals in a spoken document using an STD technique. This STD is independent of a specific STD technique. Each speech interval has one or more keywords with STD costs and the duration information.

The advantage of the SBMK-based indexing can be expected to make a more accurate index rather than a conventional ASR-based indexing method, because it is well-known that an STD engine can more precisely search a specific keyword from speech data compared with a character string search method for ASR transcription. The contribution of this paper is to demonstrate the effectiveness of the SBMK-based index on an STD task.

Fig. 1 shows the matching examples of two indexing keywords in a keyword list. 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.

Finally, the keyword (we call it "best matched keyword"), which has the STD cost under the threshold and the longest duration time among all the competitive keywords in the same interval, is selected. For all the intervals, this selection process is performed and the keywords are collected. These keywords are used for making a spoken document index. The index can be used for STD and SDR. This indexing approach can be expected to reduce false detections because one speech interval has only one keyword. This is quite different from other existing spoken document indexing methods [14]-[16] used for SDR and STD, therefore it is a radically new approach for spoken document indexing. In this paper, we experimentally demonstrate the effect of the SBMK-based indexing method on an STD task.

When the index is used for STD, first, a query is searched by an STD engine without the index in the first step, then, the search results are obtained. The second step verifies the STD results of the first step using the index and finally decides whether the first output by the STD engine is confident or not. If it is confident, the detected keyword of the first step is output as the final STD result.

We evaluate the proposed framework on academic lecture speeches as spoken documents on an STD task. It

can effectively work in annotating keyword indices to spoken documents, because the false detections are drastically reduced in the lower half of recall area in comparison of the baseline STD method. In addition, we compare with the usage of a keyword list from an oracle transcription of a spoken document in the SBMK-based indexing. The keyword list from ASR result achieves the same or better performance on the STD task.

II. Overview of the Proposed Approach

Fig. 2 shows an outline of the proposed SBMK index method of a spoken document by selecting the best matched keyword from competitive detections using an STD technique. In addition, the STD framework using the SBMK-based index also is shown at the bottom of Fig. 2.

In the SBMK-based spoken document indexing phase, first, spoken documents are performed ASR using 10 sorts of ASR systems for making a PTN, because we used the dynamic time warping (DTW)-based STD engine that uses the PTN in this paper (the details are described in Section III-A). We use DTW-based STD engine with the PTN, however, our SBMK framework is independent of specific STD methods.

Next, all the keyword in a keyword list made from the ASR system's output are searched by the STD engine. All the detected candidates are merged using the interval information of each candidate. Finally, the one keyword is selected from a competitive detection group and registered in an index of the spoken documents.

In this paper, we evaluate the SBMK-based indexing on an STD task. We use the same STD engine as the one used in the SBMK-based indexing for evaluating the STD task. When a query is input to the STD engine, first, it outputs detection candidates. They are verified using the SBMK-based index in the second step. The SBMK process in the second step decides whether the detected candidates of the query are confident or not, and finally only the confident candidates are output as the final STD result.

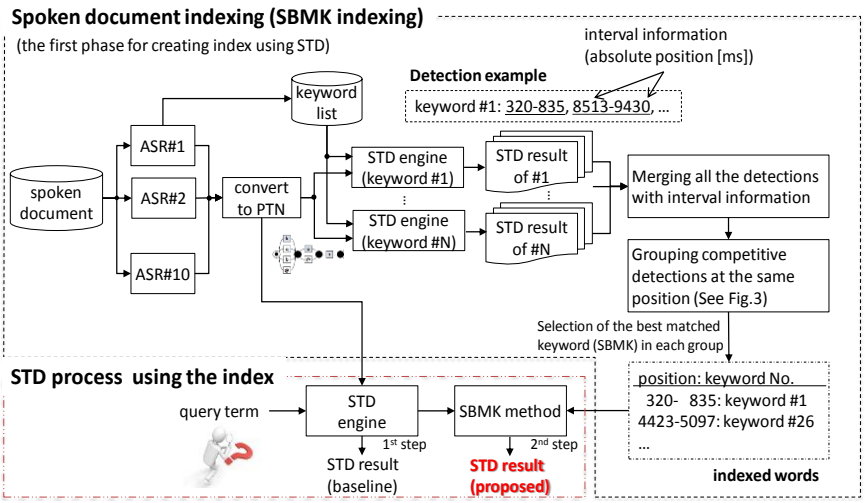

Figure 2. Workflow on indexing a spoken document using the proposed method and STD process using the index.

III. Selection of Best Matched Keyword Indexing

A. STD 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 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 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 [17], an open source decoder for ASR, is used in all systems. Acoustic models are triphone-based (Tri.) and syllable-based Hidden Markov Models (HMMs) (Syl.), both of which are trained on spoken lectures in CSJ [4]. All language models are word- and 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.

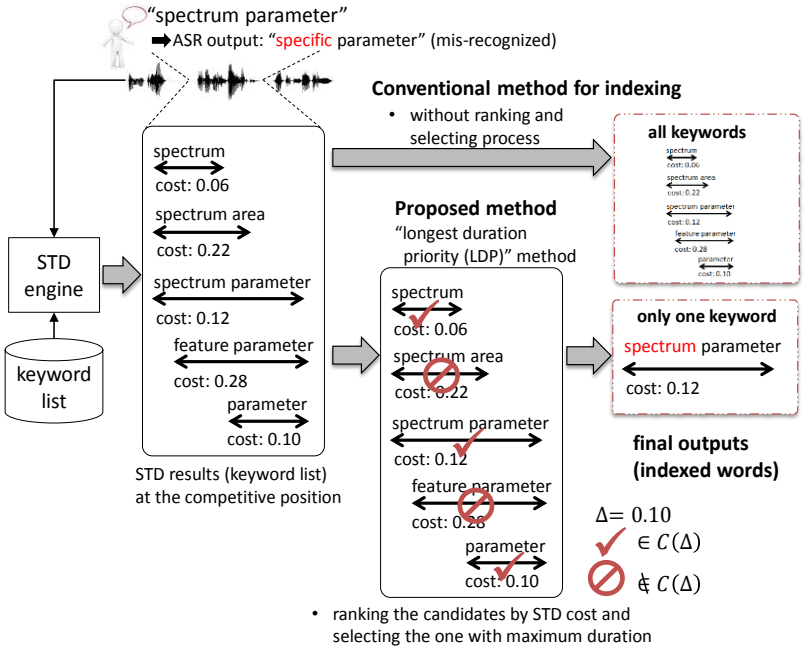

Figure 3. Keyword selection methods from a competitive interval.

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].

The word-correct and accuracy rates for the target speeches are about 68% and 63%, respectively, when the ASR system with the combination of WBC and Tri. models is used to transcribe them.

B. *Keyword List*

A keyword list is made from an ASR transcription of spoken documents by the ASR system, in which WBC and Tri. models are used. This is because the combination of these acoustic and language models achieved the best ASR performance among all the 10 ASR systems.

In the STD experiment, we will compare with three types of keyword list, which are made from following transcriptions: the transcription by the ASR system, the oracle transcription (manually transcribed), and the both transcriptions. The keyword list of the both types of transcriptions includes the keywords that exist in the path.

C. *Grouping Competitive Detections*

Spoken documents are automatically transcribed by the 10 types of ASR systems and the PTN is created from the transcription. The STD engine outputs intervals where a query term is likely to be uttered. First, we conduct STD for all the keywords in the keyword list as query terms. The set of term detections is obtained for each keyword (query term). The detection intervals of all the 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 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.

D. *Keyword Selection from Competitors*

Fig. 3 shows the two indexing methods for a spoken document. "Conventional method" is the same as a typical STD scheme that selects all the keywords belonging to C. In other words, the conventinal method outputs and indexes all keywords that have STD costs below a previously set threshold. "Proposed method", the longest-duration priority method, selects the keyword that has the longest duration among all the keywords in the competing group.

We explain the details of the proposed method below.

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:

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

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

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

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 Fig. 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:

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

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

Finally, we select the quadruple $\langle w_{ld}, t_{ld}, t'_{ld}, cost_{ld}\rangle$, 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 Fig. 3, the following quadruple is the final output:

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

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

IV. STD Using Index by SBMK Method

This section explains how to perform an STD process using the index made by the SBMK framework on the spoken document indexing phase.

The SBMK-based index by the indexing phase can be represented as a sequence of m quadruples $\langle w_{ld}, t, t', cost_{ld}\rangle$ as follows:

$$\langle w_{ld}, t, t', cost_{ld}\rangle_1, \ldots, \langle w_{ld}, t, t', cost_{ld}\rangle_m$$

These intervals $[t, t']_1, \ldots, [t, t']_m$ are not overlapped each other.

When a query keyword w_q is given to the STD engine, first the STD engine outputs term detection candidates. We define a detected candidate as a quadruple as follows and denote it as q:

$$q = \langle w_q, t_q, t'_q, cost_q\rangle$$

The interval $[t_q, t'_q]$ of $q = \langle w_q, t_q, t'_q, cost_q\rangle$ overlaps at one or more indexed keywords like below:

$$\langle w_{ld}, t, t', cost_{ld}\rangle_i, \ldots, \langle w_{ld}, t, t', cost_{ld}\rangle_j$$

In this case, we can represent the competitive detection set C as C_q below:

$$\begin{aligned} C_q = \{ &\langle w_{ld}, t, t', cost_{ld}\rangle_i, \ldots, \langle w_{ld}, t, t', cost_{ld}\rangle_j, \\ &\langle w_q, t_q, t'_q, cost_q\rangle\} \end{aligned}$$

The second step of STD verifies whether the detected candidate $q = \langle w_q, t_q, t'_q, cost_q\rangle$ is confident or not by using the SBMK framework described in Section III-D. If the STD cost cost_q of q is less than $cost_{min} + \Delta$ and q has the longest duration, q is regarded as trustworthy, then, it is outputted as the final STD result. This process is performed for all the detection candidates of the query keyword w_q.

V. Evaluation

A. Experimental Setup

We used the moderate-size STD task used in NTCIR-10 SpokenDoc-2 [6] as an STD task for evaluation. The evaluation speech data is the Corpus of Spoken Document Processing Workshop. It consists of the recordings of the first to sixth annual Spoken Document Processing Workshop, 104 real oral presentations (28.6 hours). The number of query terms is 100, where 47 of the all 100 query terms are in-vocabulary (INV) queries that are included in the ASR dictionary of the WBC LM and the others 53 queries are out-of-vocabulary (OOV).

We evaluated the two methods: the conventional method (same as "STD result (baseline)" in Fig. 2), the longest-duration priority method (proposed) on an STD task, which normally evaluates a keyword detection performance of a query set on a spoken document collection. In this paper, we make a comparison between the three methods for making a keyword set; (1) keyword set from the ASR transcription, (2) keyword set from the oracle transcription, and (3) keyword set from common words in both the ASR and the oracle transcription.

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 spoken documents as an indexing assessment because we generated keyword lists for spoken documents. An STD query set for spoken documents in the SMBK-based indexing phase was all the keywords in the keyword set, which is created by the process described in Section III-B, for the spoken document. Table I shows the number of keywords in each keyword list.

TABLE I. The Number of Keywords

Made from	# of keywords	# of indexed keywords
ASR trans.	6,575	278,197
oracle trans.	8,493	260,149
ASR and oracle trans.	4,162	284,058

The STD cost range Δ was determined as 0.20 in this paper by optimizing it using a held-out data.

B. Evaluation Metrics

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

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

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

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

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

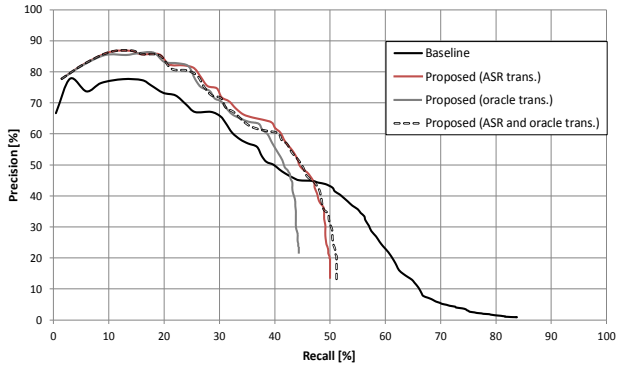

Figure 4. Recall-Precision curves for the all query set.

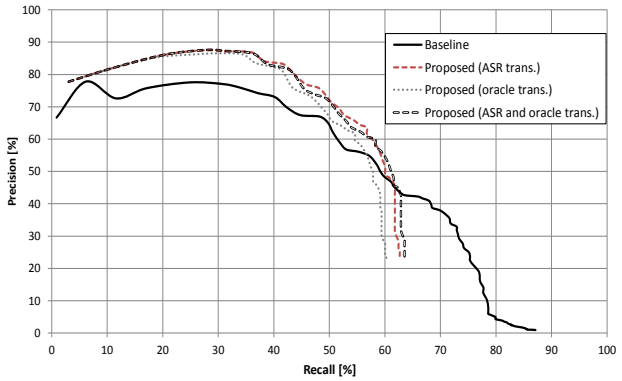

Figure 5. Recall-Precision curves for the INV query set.

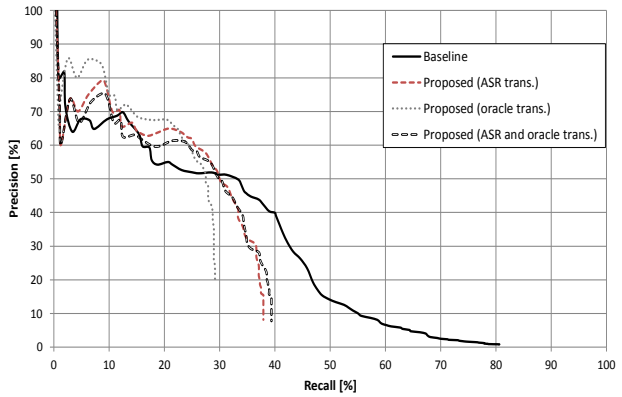

Figure 6. Recall-Precision curves for the OOV query set.

TABLE II. MAXIMUM F. VALUES FOR EACH QUERY SET [%]

		all	INV	OOV
Baseline		46.33	56.37	40.18
Proposed	ASR trans.	48.86	59.98	37.85
	oracle trans.	47.02	57.94	35.86
	ASR and oracle trans.	48.5	59.07	37.63

C. Experimental Results and Discussion

Fig. 4, Fig. 5, and Fig. 6 show the recall-precision curves for the all, the INV, and the OOV query sets, respectively. In addition, Table II also shows the maximum F-measure values for each recall-precision curve. We compare between the baseline STD (denoted as "Baseline") and three proposed SBMK-based methods that are different from the keyword lists (denoted as "Proposed (ASR trans.)", "Proposed (oracle trans.)", and "Proposed (ASR and oracle trans.)").

For all the query sets, all the proposed SBMK-based STD outperformed the precision values in the lower half of the recall range. Because our STD method can output only one keyword with the best match at the competitive intervals in which some term candidates were detected, however, it could not output correct detections at the higher recall range, where the baseline was better than the proposed methods.

Like a WEB search engine, a high-precision retrieval system is very important for fast checking the retrieved document. Therefore, our proposed framework is effective to build this retrieval system. On the other hand, the baseline STD could output more correct detections (over 80% recall rate) because it is robust for ASR errors and OOV terms. Therefore, the combination of the baseline and the proposed method is sure to drastically improve the whole of recall-precision performances.

Comparing with the three proposed methods (three types of keyword lists), "ASR trans." got the best STD performance on the INV query set, however, it had a weakness for the OOV query set compared to "oracle trans." because it included the OOV query terms. On the one hand, as shown in Fig. 5, the whole of recall-precision curve of "oracle trans." was worse than "ASR trans." and "ASR and oracle trans." in the whole performance of recall-precision curve for the INV query set. This may be because the keywords that are included in the oracle transcription but are not included in the ASR transcription damaged the STD performance. Most of these keywords are difficult to correctly speech-recognize, and they are also not important (relevant) to STD query. In addition, because these keywords had more number of phonemes rather than the input query terms, they were likely to beat the other competitors at the competitive intervals.

From the above results, it is shown that the keyword list form the ASR transcription worked better for the INV query set rather than the oracle transcription. It is very effective for practical use because the oracle transcription cannot used in the STD system. Therefore, it can be said that our SBMK-based indexer for spoken document contributed to improve the STD performance.

VI. CONCLUSION

This paper described a novel spoken document indexing method using keywords collected from the ASR result of the target document. Our indexing method uses an STD engine to indicate speech intervals of a word included in the keyword list. The one keyword (best matched keyword) is selected based on the STD cost and the duration time of the keyword from the competitive interval in which one or more keywords are attached. The selected keywords are registered into an index. We evaluated our indexing method on the STD task of NTCIR-10 SpokenDoc-2. The experimental result of the STD task showed that our technique achieved great

reduction of the false detections at the lower half of the recall rate for the INV query set despite using the keyword list made from the errorful transcription. In future studies, we are going to introduce a machine learning framework such as support vector machines for selecting one keyword from a competitive interval.

ACKNOWLEDGMENTS

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 *Proc. TREC8*, 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 *Proc. 11th INTERSPEECH*, 2010, pp. 677-680.
[3] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, 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, 2009.
[4] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1-8.
[5] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for spoken documents task in NTCIR-9workshop," in *Proc. 9th NTCIR*, 2011, pp. 223-235.
[6] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamashita, "Overview of the NTCIR-10 SpokenDoc-2 task," in *Proc. 10th NTCIR*, 2013, pp. 573-587.
[7] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones, "Overview of theNTCIR-11 SpokenQuery&Doc task," in *Proc. 11th NTCIR*, 2014, pp. 350-364.
[8] The Spoken Term Detection (STD) 2006 evaluation plan. (2006). [Online]. Available: http:www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.%pdf
[9] K. Iwata, K. Shinoda, and S. Furui, "Robust spoken term detection using combination of phone-based and word-based recognition," in *Proc. 9th INTERSPEECH*, 2008, pp. 2195-2198.
[10] B. Logan and J. V. Thong, "Confusion-Based query expansion for OOV words in spoken document retrieval," in *Proc. 7th ICSLP*, 2002, pp. 1997-2000.
[11] 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 *Proc. 8th INTERSPEECH*, 2007, pp. 2393-2396.
[12] 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 *Proc. 9th INTERSPEECH*, 2008, pp. 2146-2149.
[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] M. Kurimo, "Fast latent semantic indexing of spoken documents by using self-organizing maps," in *Proc. ICASSP*, 2000, vol. 4, pp. 2425-2428.
[15] Y. Pan, H. Chang, and L. Lee, "Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing," in *Proc. ASRU*, 2007, pp. 677-682.
[16] F. Seide, K. Thambiratnam, and P. Yu, "Word-Lattice based spoken document indexing with standard text indexers," in *Proc. SLT*, 2008, pp. 293-296.
[17] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. 1st APSIPA ASC*, 2009, pp. 1-6.

Kentaro Domoto received his B.E. degree in engineering from University of Tsukuba in 2014. He is now a student of master course in Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba. His research interests include spoken language processing, in particular speech term detection. He is a student member of the Acoustical Society of Japan.

Takehito Utsuro received his B.E., M.E., and D.Eng. degrees in electrical engineering from Kyoto University in 1989, 1991, and 1994. He has been a professor at the Division of Intelligent Interaction Technologies, Faculty of Engineering, Information and Systems, University of Tsukuba, since 2012. His professional interests are in natural language processing, Web intelligence, information retrieval, machine learning, spoken language processing, and artificial intelligence. He is a member of the Association for Computational Linguistics (ACL) the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), the Japan Society for Artificial Intelligence (JSAI), and the Acoustical Society of Japan.

Naoki Sawada received his B.E. degree in computer and media sciences from University of Yamanashi in 2014. He is now a master course student at Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi. His research interests include spoken language processing, in particular speech term detection. He is a student member of the Acoustical Society of Japan.

Hiromitsu Nishizaki received his B.E., M.E., and D. Eng. degrees in information and computer sciences from Toyohashi University of Technology in 1998, 2000, and 2003. He is now an assistant professor in Department of Research, Interdisciplinary Graduate School of Engineering at University of Yamanashi. His research interests include spoken language processing, speech interface, and human-computer interaction. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), and the Acoustical Society of Japan.