Open Vocabulary Spoken Content Retrieval by front-ending with Spoken Term Detection

Tomoko Takigami* and Tomoyosi Akiba*
*Department of Computer Science and Engineering, Toyohashi University of technology, Japan
E-mail: akiba@cs.tut.ac.jp

Abstract—**How to deal with speech recognition errors and out-of-vocabulary (OOV) words are common challenging problems in spoken document processing. In this work, we propose the spoken content retrieval (SCR) method that incorporates spoken term detection (STD) as the pre-processing stage. The proposed method firstly performs STD for each term appearing in the given query topic, then the detection results are used to calculate the relevance of the retrieved document to the topic. The front-ending with STD enables to make use of even misrecognized and OOV words as the clues of the back-end document retrieval process. We also propose a novel retrieval model especially designed for the proposed SCR method. It incorporates the term co-occurrences into the conventional vector space model in order to put emphasis on reliable clues for the similarity calculation, which enables the retrieval process to work robust for documents including errors. The experimental results showed that the performance of the proposed SCR method improved the retrieval performance when a query topic included OOV words, even though it relied on the lower-accuracy syllable-based ASR results. They also showed that the proposed retrieval model significantly improved the retrieval accuracy not only for the proposed SCR but also for the conventional SCR methods.**

I. Introduction

In recent years, multimedia contents have exploded and search technologies for those contents have become crucial. However, much of those contents do not have metadata. Therefore, searching such contents using only traditional text-based methods is difficult. On the other hand, contents that include speech can be searched using its language information with the help of large vocabulary continuous speech recognition (LVCSR). Search technology that targets spoken language information is called spoken document retrieval (SDR). Unlike text retrieval, dealing with speech recognition errors and out-of-vocabulary (OOV) words are challenging problems in spoken document processing.

Among retrieval tasks involving spoken documents, the task of finding the occurrences of a given term of interest within the speech data is called spoken term detection (STD)[1][1]. Many studies of STD have focused on dealing with OOV and misrecognized words. Such methods include, for example, using subword units as the recognition result or search unit[5], [6], [7], using the multiple recognition candidates [5], [6], using the results of multiple recognition systems[8], and using soft matching to detect the occurrences[9], [10].

Another retrieval task targeting spoken documents is called spoken content retrieval (SCR). The task aims to find a document that is relevant to a given query topic, where the relevancy is judged by a human assessor[2]. The conventional SCR approach simply applies text-based retrieval to the recognition results obtained by LVCSR. In SCR, a query topic includes multiple terms, each of which can be used as a clue for the retrieval; thus, SCR can be robust against OOV and misrecognized words. In fact, in the TREC SDR evaluation, the conventional approach for SCR achieved retrieval accuracies almost equal to those using human transcription, and ad hoc retrieval for broadcast news was considered a "solved problem" [11]. However, in real-life applications, high recognition rates cannot always be expected. Therefore, OOV and misrecognized words should be dealt with even in the SCR task.

The effect of misrecognition can be relaxed by using the multiple recognition candidates of speech recognition results [12]. On the other hand, OOV words can never be handled if we rely only on word-based recognition results of spoken documents and their word indices obtained from them. For dealing with OOV words, subword-based recognition results and their subword n-gram indices have been often used in spoken document retrieval [13], [14], [15]. However, as the discriminative power of these subword n-grams is much weeker than that of the whole words, the performance of the SCR system based on such subword n-grams is limited.

In this work, we propose a method that incorporates STD into the SCR process to deal with OOV and misrecognized words. In the first step, an STD method is applied to the spoken documents, where each term in the given query topic is searched for in the syllable sequence obtained by speech recognition. From the detection results, we can obtain the statistics for term frequencies in each document, to which we can apply any conventional document retrieval method. The advantage of this method is that it is not affected by the OOV terms in the query topics even though it still makes use of the whole word clues. The proposed method resembles the early approach in SCR [16], [17], which applies word-spotting for spoken documents instead of LVCSR. As the word-spotting had to be applied after a query topic was given,

[1] Note that the SDR task that aims to find a longer segment, e.g. a document, that includes given terms can be considered rather as a variation of an STD task than as an SCR task, because the success of the search task depends only on the occurrences of the terms in a document. The works on such a task include [2],[3], [4], etc. It was called "known-item retrieval" in the TREC SDR track.

[2] The task was called "Ad-hoc retrieval" in the TREC SDR track.

it was not a tractable approach for targetting a large amount of spoken documents. Thanks to the recent development of the fast STD methods, the approach now become tractable even for targetting a large amount of spoken documents, which is investigated in this work.

We also propose a novel retrieval model especially designed for text including errors by extending conventional vector space model for text retrieval. The proposed method is unique in using the term co-occurrences in order to put emphasis on reliable clues in transcribed text for the similarity calculation, which results in working robust for documents including errors.

This paper is organized as follows. The next section describes our SCR methods. In Section 3, we describe our proposed retrieval model. In Section 4, we evaluate the proposed method by comparison with conventional document retrieval methods. Finally, we conclude and describe future work in Section 5.

II. Spoken Content Retrieval

A. Conventional Method

The conventional SCR system works as follows. First, we create automatic transcripts by applying LVCSR to speech data. Then, we convert the transcripts into a bag-of-words representation by applying the word segmenter, the lemmatizer, and stop-word removal to the text. Indexing can also be applied for efficient retrieval. Given a query topic, it is converted similarly into a bag-of-words. Then, any conventional information retrieval method can be used to calculate the similarity between these bag-of-words representations. For example, in the vector space model, the bag-of-words representations of the documents and the query topic are converted into a vector representation, and then the inner product between the vectors is calculated to find documents similar to the query topic.

The conventional SCR methods use word-based speech recognition to obtain the transcription of the spoken documents, and then text-based document retrieval is applied to the transcription. However, the OOV words from the word-based speech recognition and misrecognized words can never be used as the clues for the document retrieval, which results in the degradation of the retrieval performance. To deal with these problems, both document and query expansion methods have been proposed [18], [19], [20]. These methods are not using OOV words or speech segments of recognition errors in the document, but these extensions of related words as clue words that are correctly recognized for order to deal with recognition errors. On the other hands, subword n-gram based SCR methods have been proposed [13], [14], [15] . While the conventional SCR used words as a feature of a document vector, these methods used subword n-gram. However, subword n-grams are a automatically extracted without using the language knowledge, the effectiveness of such clues is limited compared to the word.

B. STD-based SCR

To deal with the problem of recognition errors and OOV words, we propose the STD-based approach for SCR. Figure 1 shows the configuration of our STD-based system.

First, we create automatic transcripts by applying subword-based recognition to speech data. Given query topic, the keywords are extracted from the query topics and are converted to subword sequences. In this paper, we use nouns as the keywords and syllables as the subword unit. Then, each syllable sequence is used as a search key against the syllable-based transcription of the spoken documents. From the STD results, we can obtain the keyword frequency for each document in the collection. We repeat the process for all keywords, and then we can obtain the vector of the keyword frequencies for each document. Finally, the vectors are compared with the query vector to select the retrieval results, which is equivalent to applying the conventional SCR system.

Note that the proposed method is different from the previous works that use the subword n-grams as clues [13], [14], [15]. While they also use the subword-based recognition results, their document and query vectors are created based on the subword n-grams. Or rather, our proposed method resembles the early approach in SCR [16], [17], which applies word-spotting for spoken documents instead of LVCSR. As the word-spotting had to be applied after a query topic was given, it was not a tractable approach for targetting a large amount of spoken documents. Thanks to the recent development of the fast STD methods, the approach now become tractable even for targetting a large amount of spoken documents, which is investigated in this work.

1) STD method: The method as STD was used Dynamic Time Warping (DTW) algorithm according to the following equation.

$$\begin{aligned} D_{0,j} &= 0\ (0 \leq j \leq J) \\ D_{i,0} &= \infty\ (1 \leq i \leq I) \\ D_{i,j} &= \min\{D_{i,j-1}, D_{i-1,j-1}, D_{i-1,j}\} + d(a_i, b_j), \end{aligned} \quad (1)$$

where i is the position in the keyword, and j is the position in the spoken documents. $d(a_i, b_j)$ is a distance between syllables a_i and b_j, and $D_{i,j}$ is a cumulative distance. The cumulative distances at the tail of the keyword is normalized by its length and the detection is made if the normalized distance is below some predefined threshold.

However, DTW algorithm is time consuming as it must traverse all the target spoken documents. Especially in our SCR setting, as a query topic contains multiple terms, multiple DTW runs must be executed.

Therefore, we tried to improve its time efficiency for STD by skipping the utterances that is less likely to have the queried keyword.

First, the syllable bi-gram index is created from the spoken documents. For each utterance U, which is a syllable sequence of a text $U = t_1 t_2 \ldots t_n$, we create syllable bi-gram index B_U

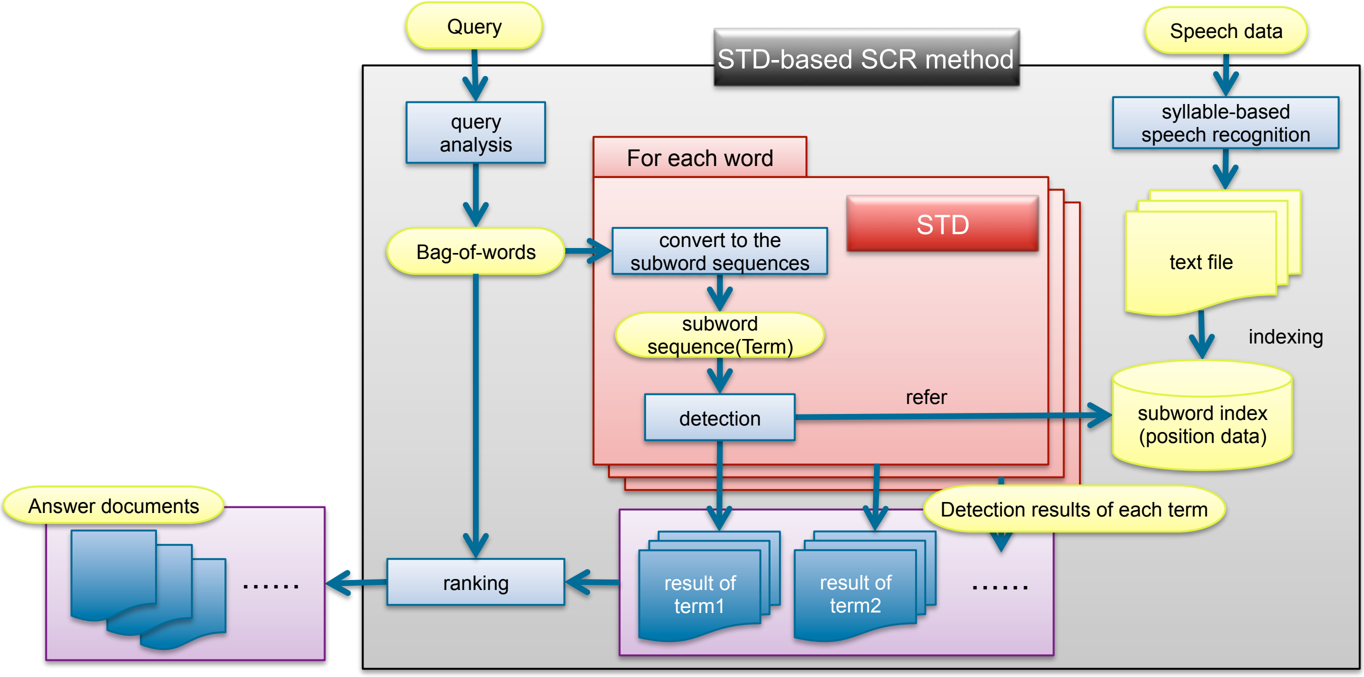

Fig. 1. *STD-SCR system*

as follows.

$$B_U = \{(t_i, t_{i+1}) \mid i = 1, 2, \ldots, n-1\} \quad (2)$$

Here, the inverted file is used to record the document ID and the syllable ID for each bi-gram indexes.

At the search time, syllable bi-gram index B_Q of a query term $Q = w_1 w_2 \ldots w_m$ is created, too.

$$B_Q = \{(w_i, w_{i+1}) \mid i = 1, 2, \ldots, m-1\} \quad (3)$$

Then, DTW algorithm is performed only for the utterances that satisfy the following equation.

$$\frac{\sum_{b \in B_Q} \delta(b \in B_U)}{|B_Q|} \geq \theta \quad (4)$$

$$\delta(x \in X) = \begin{cases} 1 & (x \in X) \\ 0 & (\text{otherwise}) \end{cases}, \quad (5)$$

where θ is a predetermined threshold. In this paper, we fix θ in a constant regardless of the threshold of DTW algorithm.

C. Combination of CSCR and STD-SCR

The proposed method will be effective for the query including the words that are OOV or misrecognized in the spoken documents, while the conventional SCR will be effective for the query that consists of the words recognized correctly in the spoken documents. It is expected that the two systems can complement each other, so it is worth investigating the hybrid system of them.

Not only simply combining the both systems, we tried to further boost the performance by making a distinction between OOV and IV words in a given query topic. Firstly, we divide the words in a query topic q into the OOV words q_{OOV} and IV words q_{IV} by consulting the recognition dictionary of the LVCSR system used in the conventional SCR system. Then, we combine the two systems as follows.

$$\begin{aligned} sim(q, d) \;=\; & (1-\alpha)\, sim_{cSCR}(q, d) \\ & +\alpha\{(1-\beta)\, sim_{STD-SCR}(q_{IV}, d) \\ & +\beta sim_{STD-SCR}(q_{OOV}, d)\}, \end{aligned} \quad (6)$$

where, α and β are weighting coefficients of the linear combination, sim_{cSCR} and $sim_{STD-SCR}$ are relevance scores calculated in the conventional SCR system and the proposed STD-SCR system, respectively. Using higher α means that we prefer the STD-SCR system to the conventional SCR system. Note that $\alpha = 0$ ($\alpha = 1$) corresponds to just using only the conventional SCR (STD-SCR) system. On the other hand, using higher β means that OOV words are taken more importance than IV words when using the STD-SCR system. Note that $\beta = 0.5$ means that we make no distinction between OOV and IV words.

III. Retrieval Model for SCR

As a method to calculate the similarity between each document and a given query topic, vector space model has been widely used in document retrieval. Unlike text retrieval, as the occurrence frequency of each word in SCR is uncertain caused by recognition errors, it can not be used as a reliable clue for document retrieval. In this work, we propose a novel retrieval model that extends the vector space model robust for uncertain clues.

There are two types of errors brought in the document vector obtained by using automatic speech recognition. One is so called *false negative*, which is an error of the word that appears in the document but not in the recognition result. The other is so called *false positive*, which is an error of the word that does not appear in the document but in the recognition result. Previous works on SCR have mainly focused on the false negative. For example, query expansion [18], [19] and document expansion [20] have been well studied in the context of SCR, which aim to compensate for the false negative by introducing the other words than the misrecognized words.

On the other hand, there have been few works that aim to compensate for the false positive. Actually, the false positive appears much less than the false negative within the conventional SCR based on the LVCSR result, so it does not affect much. However, our proposed STD-SCR tends to have many false positive as it is based on the STD detection results. In this work, we also propose the novel retrieval model designed for our proposed STD-SCR.

A. Vector space model using word combination feature

Two keywords contained in a query topic are related to each other, therefore, we can be considered these words are more likely to appear at the same time in the document. We attempt to make use of word co-occurrence as an additional feature that is used in computing the similarity between the query and each document.

For document d, document vector $v(d)$ is computed as follow,

$$v(d) \quad = \quad [tf(w_1, d), tf(w_2, d), \dots], \tag{7}$$

where $tf(w, d)$ is frequency of word w in document d. Additionally, we also calculate co-occurrence information vector $v_c(d)$ as follows.

$$v_c(d) \quad = \quad [\delta(w_1, w_2, d), \delta(w_1, w_3, d), \dots, \delta(w_i, w_j, d), \dots] \tag{8}$$

$$\delta(w_i, w_j, d) \quad = \quad \begin{cases} 1 & (w_i \in d \text{ and } w_j \in d) \\ 0 & (\text{otherwise}) \end{cases}, \tag{9}$$

The size of the co-occurrence vector $v_c(d)$ is $(|v(d)|^2 - |v(d)|)/2$. Figure 2 shows example of document vector for query $q = w_1, w_2, w_3$.

Given query topic q, firstly, we obtain document vector $v(d)$ using word frequency that contain query topic. Than, co-occurrence vector $v_c(d)$ is computed based on $v(d)$, and extended vector $v_e(d) = [v(d), v_c(d)]$ is obtained by concatenating the $v(d)$ and $v_c(d)$. Similarly, the extended query vector $v_e(q)$ is also calculated. Finally, the similarity between $v_e(d)$ and $v_e(q)$ is calculated as same as the vector space model with TF-IDF term weighting.

	$v(d)$			$v_c(d)$		
	w_1	w_2	w_3	w_1w_2	w_1w_3	w_2w_3
d_1	3	0	1	0	1	0
d_2	1	2	2	1	1	1
d_3	0	0	1	0	0	0

Fig. 2. *Example of document vector by extended vector space model*

IV. Experiments

A. Experimental Setting

We used the SCR test collection developed in the NTCIR-9 Workshop [21] for our experiments. The target document collection contains 2702 lectures selected from the Corpus of Spontaneous Japanese (CSJ)[22]. This amounts to more than 600 hours of speech.

The test collection contains two set of queries, one contains 39 queries called dry-run set, another one contains 86 queries called formal-run set. Originally, these queries required to find a passage of varying lengths from lectures. However, in this paper, we were set to pseudo-passage unit as retrieval unit, as same as the task definition often used in the literature [23], [24]. We defined pseudo-passages by automatically segmenting each lecture into sequences of segments with fixed numbers of sequential utterances. We used 15 utterances in a segment, which results in the number of pseudo-passages found in the target document to be 60,375. Next, we assigned retrieved pseudopassages a relevance label as follows: if the pseudopassage shared at least one utterance that came from the relevant passage specified in the "golden file," then the pseudopassage was labeled as "relevant."

We constructed two ASR systems, an LVCSR (word-based model) and a continuous syllable recognition system (syllable-based model) for recognition of spoken documents. The language models of both systems were trained by using newspaper articles covering 75 months and we used word trigrams for the word-based model and syllable trigrams for the syllable-based model. The acoustic model was trained by using the CSJ data by keeping the open condition[23], and we used the same model for both systems. The word accuracy rate of LVCSR was 59.5%, and the syllable accuracy rates of the two systems were 80.6% and 75.5%, respectively.

We used the indexing and DTW algorithm described in Section II-B1 for the STD part of the proposed STD-SCR method. The STD was performed on the result by the syllable-based ASR system. For the distance between two syllables for DTW algorithm, we used Battacharyya distance between the acoustic models. The detection threshold was determined through the 2-fold cross-validation on the experimental data. For the document retrieval, we used the vector space model with TF-IDF term weighting.

We compared our STD-based approach with the conventional SCR approach as a baseline (cSCR). The baseline system used the word-based automatic transcription as a textual representation of the spoken documents and applied the conventional vector space model with TF-IDF term weighting

TABLE I
System setting

	cSCR-word	cSCR-subword	STD-SCR	hybrid SCR
ASR	word-based		syllable-based	word-based & syllable-based
indexing	word	syllable bi-gram	syllable bi-gram	word & syllable bi-gram
document retrieval	vector space model with TF-IDF term weighting			

for the document retrieval, which was the same as those used in the SCR part of our STD-based approach. As the indexing unit for cSCR, either word (*cSCR-word*) or syllable bi-gram (*cSCR-subword*) [3] was used.

In addition, we also investigated the performance of the hybrid system that is a mixture of *STD-SCR* with *cSCR-word* (*hybrid SCR*) described in section II-C. Table I shows the conditions of the experimental systems.

B. Experimental results

Table II shows experimental results using the conventional vector space model, where we use mean average precision (MAP) as the evaluation metric. The row labeled ALL corresponds to the MAP averaged over all the query topics, while the row labeled OOV (or IV) corresponds to the MAP averaged over only those including at least one OOV term (or only IV terms). The numbers of OOV and IV queries are 41 and 84, respectively.

For ALL queries, *cSCR-word* performed better than *cSCR-subword* and *STD-SCR*. However, when looking at the results for OOV queries, while the retrieval performances of both the conventional methods, i.e. *cSCR-word* and *cSCR-subword*, largely degraded, the performance of proposed *STD-SCR* was stable. It indicated the robustness of *STD-SCR* for OOV queries. Furthermore, *hybrid SCR* performed best among the best for all types of the queries.

Table III shows experimental results using the proposed vector space model using word co-occurrence features. It indicated that the proposed retrieval model significantly improved the retrieval accuracy not only for *STD-SCR* but also for the conventional SCR methods.

Figure 3 shows the relationship between STD performance (F-measure) and SCR performance. It seems that STD and SCR performances have strong positively correlation. It indicated that the improvement on the STD performance would resulted in the improvement on the SCR performance for our proposed STD-SCR method.

V. Conclusion

In this paper, we propose a method that incorporates STD into SCR process to deal with OOV and miss-recognized words. Through our experimental evaluation, we found that the performance of the proposed STD-SCR improved the retrieval performance when a query topic included OOV words, even though it relied on the lower-accuracy syllable-based ASR results. We also found that the proposed retrieval model significantly improved the retrieval accuracy not only for the STD-SCR but also for the conventional SCR methods. Furthermore, we found that there is strong positive correlation between STD and SCR performances. This indicates that SCR performance will be improved by improving the performance of the STD. In future work, we are going to introduce query expansion into the proposed STD-SCR framework in order to further improve the performance of SCR.

References

[1] National Institute of Standards and Technology, "Spoken term detection evaluation portal," "`http://www.nist.gov/speech/tests/std/`".

[2] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2005, pp. 443–450.

[3] H. Nishizaki and S. Nakagawa, "Robust spoken document retrieval methods for misrecognition and out-of-vocabulary keywords," *Systems and Computers in Japan*, vol. 35, no. 14, pp. 44–52, 2004.

[4] J. Fayolle, M. Saraçlar, F. Moreau, C. Raymond, and G. Gravier, "Lexical-phonetic automata for spoken utterance indexing and retrieval," in *Proceedings of International Conference on Speech Communication and Technology*, 2012.

[5] M. Saraçlar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proceedings of Human Language Technology Conference*, 2004.

[6] P. Yu and F. Seide, "A hybrid word / phoneme-based approach for improved vocabulary-independent search in spontaneous speech," in *Proceedings of International Conference on Spoken Language Processing*, 2004.

[7] Y. Itoh, K. Iwata, M. Ishigame, K. Tanaka, and L. Shi-wook, "Spoken term detection results using plural subword models by estimating detection performance for each query," in *Proceedings of International Conference on Speech Communication and Technology*, 2011, pp. 2117–2120.

[8] S. Natori, H. Nishizaki, and Y. Sekiguchi, "Japanese spoken term detection using syllable transition network derived from multiple speech recognizers' outputs," in *Proceedings of International Conference on Speech Communication and Technology*, 2010, pp. 681–684.

[9] D. Wang, S. King, J. Frankel, and P. Bell, "Stochastic pronunciation modelling and soft match for out-of-vocabulary spoken term detection," in *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, 2010, pp. 5294–5297.

[10] K. Katsurada, S. Sawada, S. Teshima, Y. Iribe, and T. Nitta, "Evaluation of fast spoken term detection using a suffix array," in *Proceedings of International Conference on Speech Communication and Technology*, 2011, pp. 909–912.

[11] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of TREC-9*, 1999, pp. 107–129.

[12] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "A lattice-based approach to query-by-example spoken document retrieval," in *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, 2008, pp. 363–370.

[13] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Communication*, vol. 32, no. 3, pp. 157–186, 2000.

[14] B. Chen, H. min Wang, and L. shan Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in mandarin chinese," *IEEE Transactions on Speeh and Audio Processing*, vol. 10, pp. 303–314, 2002.

[3] We also investigated the variant of the *cSCR-subword* that used the syllable-based automatic transcription instead of the word-based one, but we found it performed worse.

TABLE II
Retrieval performances using conventional vector space model (MAP)

	cSCR-word	cSCR-subword	STD-SCR	hybrid SCR
ALL	**0.0927**	0.0659	0.0744	0.107
IV	**0.112**	0.0772	0.0795	0.123
OOV	0.0468	0.0421	**0.0639**	0.0743

TABLE III
Retrieval performances using vector space model with word co-occurrences (MAP)

	cSCR-word	cSCR-subword	STD-SCR	hybrid SCR
ALL	**0.111**	0.0919	0.0845	0.129
IV	**0.138**	0.104	0.0885	0.148
OOV	0.0452	0.0659	**0.0762**	0.0893

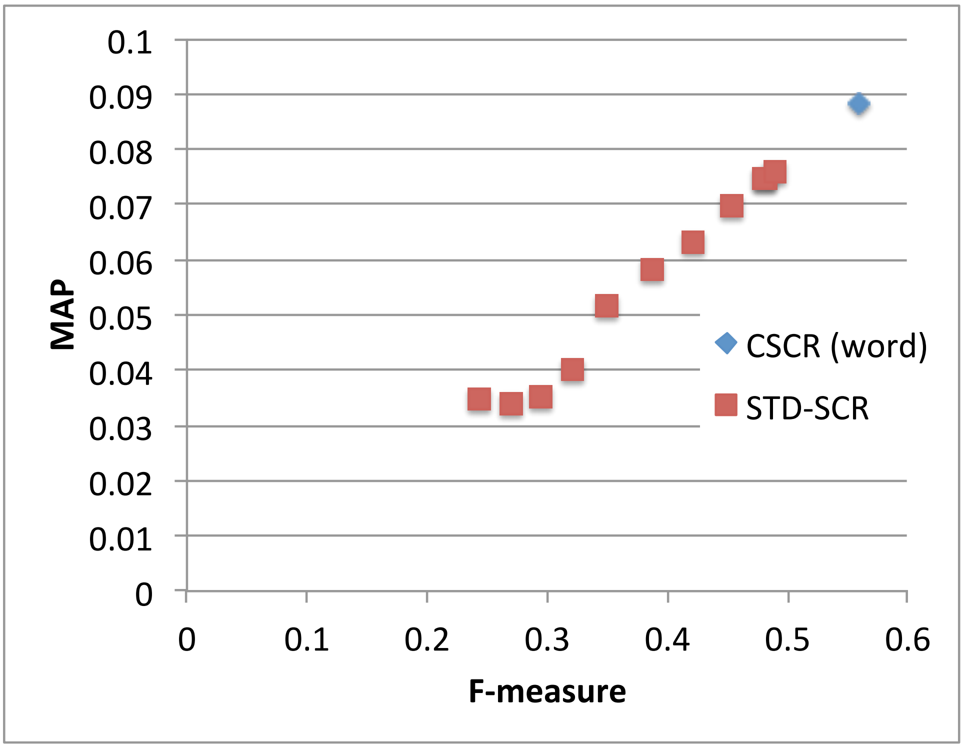

Fig. 3. *The relationship between STD performance and SCR performance*

[15] Y.-c. Pan and L.-s. Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 6, pp. 1562–1574, Aug. 2010.

[16] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," 1996.

[17] M. Wechsler, E. Munteanu, and P. Schäuble, "New techniques for open-vocabulary spoken document retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 20–27.

[18] T. Akiba and K. Honda, "Effects of query expansion for spoken documnet passage retrieval," in *Proceedings of International Conference on Speech Communication and Technology*, 2011, pp. 2137–2140.

[19] M. Terao, T. Koshinaka, S. Ando, R. Isotani, and A. Okumura, "Open-vocabulary spoken-document retrieval based on query expansion using related web documents," in *Proceedings of International Conference on Speech Communication and Technology*, 2008, pp. 2171–2174.

[20] K. Sugimoto, H. Nishizaki, and Y. Sekiguchi, "Effect of document expansion using web documents for spoken documents retrieval," in *Proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2010, pp. 526–529.

[21] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Designing an evaluation framework for spoken term detection and spoken document retrieval at the NTCIR-9 SpokenDoc task," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[22] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of International Conference on Language Resources and Evaluation*, 2000, pp. 947–952.

[23] T. Akiba, K. Aikawa, Y. Itoh, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, and K. Itou, "Developing an sdr test collection from japanese lecture audio data," in *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, 2009, pp. 324–330.

[24] H. Nanjo, Y. Iyonaga, and T. Yoshimi, "Spoken document retrieval for oral presentations integrating global document similarities into local document similarities," in *Proceedings of International Conference on Speech Communication and Technology*, 2010, pp. 1285–1288.