High Priority in Highly Ranked Documents in Spoken Term Detection

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

***Abstract*— In spoken term detection, the retrieval of OOV (Out-Of-Vocabulary) query terms are very important because query terms are likely to be OOV terms. To improve the retrieval performance for OOV query terms, the paper proposes a re-scoring method after determining the candidate segments. Each candidate segment has a matching score and a segment number. Because highly ranked candidate is usually reliable and a user is assumed to select query terms so that they are the special terms for the target documents and they appear frequently in the target documents, we give a high priority to the candidate segments that are included in highly ranked documents by adjusting the matching score. We conducted the performance evaluation experiments for the proposed method using open test collections for SpokenDoc-2 in NTCIR-10. Results showed the retrieval performance was more than 7.0 points improved by the proposed method for two test sets in the test collections, and demonstrated the effectiveness of the proposed method.**

I. INTRODUCTION

The increase of the capacity of the recording mediums such as a hard disk or an optics disk in these years enables every user to deal with multimedia data that are available on such hard disk video recorders or the Internet. Researches of SDR (Spoken Document Retrieval) and STD (Spoken Term Detection) have been conducted among speech processing researchers to deal with such enormous video data as spoken documents [1]-[3]. A common STD system generates a transcription of speech data using a LVCSR (large vocabulary continuous speech recognition) system for finding IV (In-Vocabulary) query terms at high speed, and a subword recognition system for detecting OOV (Out-Of-Vocabulary) query terms that are not included in a dictionary of the LVCSR system. Because query terms are likely to be OOV terms, such as technical terms, geographical names, personal names and neologism and so on, STD systems must be able to detect OOV query terms. The detection of OOV query terms is realized by using subword such as monophone and triphone [4][5]. The subword based system compares a query subword sequence with all of the subword sequences in the spoken documents and retrieves the target segments using CDP (Continuous Dynamic Programming) algorithm. Each candidate segment has a matching score (CDP score) and a segment number. The paper proposes a re-scoring method to improve the retrieval performance after determining the candidate segments by CDP scores. We give a high priority to the candidate segments that are included in highly ranked documents by adjusting CDP score. The basic idea for the proposed method is that a highly ranked candidate is usually reliable and a user selects query terms so that they are the special terms for the target documents and appear frequently in the target documents. Therefore, we give a benefit to the score of a candidate segment that is included in the document that already appears in a higher ranked candidate.

The research [6][7] improved the STD performance by re-ranking candidate segments through computing acoustic score in detail in the latter stage. The STD performance was improved by applying highly ranked candidates to sophisticate the acoustic score directly [8]. The proposed method utilizes the document information that includes highly ranked candidates rather than acoustic information of highly ranked candidates.

The present paper describes the outline of our system first, and subword based STD process using subword models and phonetic distances for a local distance of CDP. In Chapter 3, the proposed method giving a high priority to the candidate segments that are included in highly ranked documents is described in detail. In Chapter 4, the performance of the proposed method is evaluated using test collection for SpokenDoc-2 of NTCIR-10 [9]. Conclusions are presented lastly.

II. OUR STD SYSTEM FOR OOV QUERY TERMS

The outline of our STD system is shown in Figure 1. In our STD system for OOV query terms [10][11], first, subword recognition is performed for all of the spoken documents and subword sequences of spoken documents are prepared beforehand (1) using subword acoustic models, their language models, and a subword distance matrix. Here, subword language models are used, such as subword bigrams and trigrams and so on. The system allows both text and speech queries (2). When a user inputs a text query, the text is automatically converted to a subword sequence according to conversion rules (3). In case of Japanese, the phone sequence to be pronounced of a query term is automatically obtained when a user input a query term.

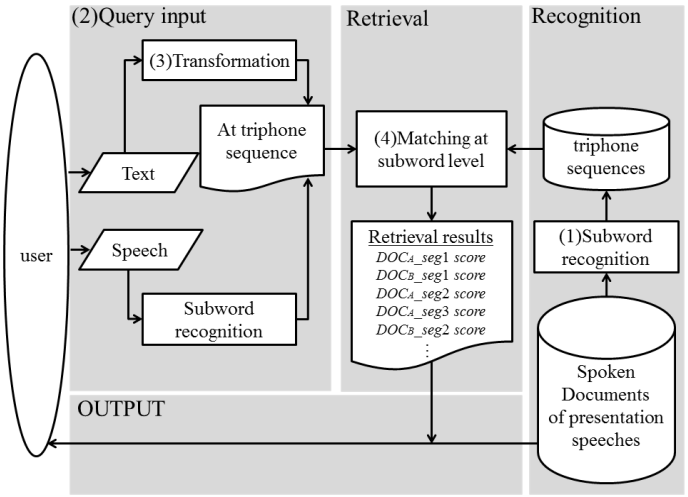


Fig. 1 Outline of the STD method using subword recognition results.

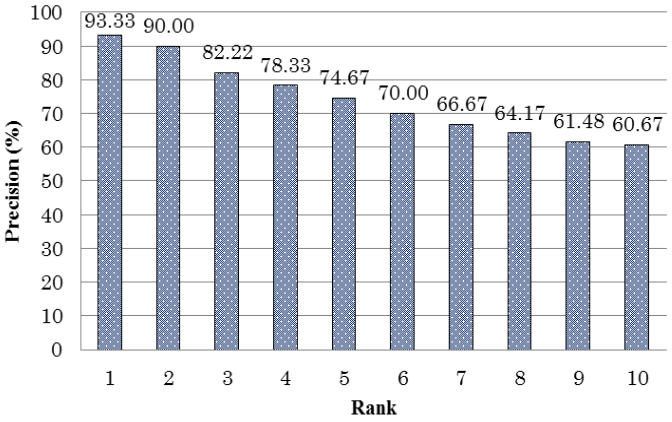


Fig. 2 Precision rates from the top candidate to among 10 candidates (the average for 30 query terms) (30 query)

For speech queries, the system performs subword recognition and transforms the speech into a subword sequence in the same manner as spoken documents (4). For each subword model, the system then retrieves the candidate segment using CDP algorithms by comparing a query subword sequence to all of the subword sequences in the spoken documents. The local distance refers to the distance matrix that represents the subword dissimilarity and contains the statistical distance between any two subword models. Although an edit distance is representative for a local distance in string matching, we have proposed a phonetic distance between subwords so far [12]. The system output candidate segments that show a high degree of similarity to the query word. Each candidate segment has a distance (CDP score) and a segment number of spoken documents. The rank of candidate segments is determined according to the CDP scores.

III. PROPOSED METHOD

This chapter describes the proposed method giving a high priority to the candidate segments that are included in highly ranked documents is described in detail.

In generally speaking in STD, highly ranked candidate segments are reliable, as seen in a high precision rate for top candidates. It is assumed that a user selects query terms so that they are the special terms for the target documents and appear frequently in the target documents. Actually the selected query terms can be expected to appear frequently in the target documents. Therefore, we give a high priority to the candidate segments that are included in highly ranked documents. We give a benefit to the CDP score of a candidate segment that is included in the document that already appears in a higher ranked candidate. We believe the method enables correct but low ranked candidates because of subword recognition errors to rank higher and improve the STD performance.

A. Analysis of highly ranked candidates and occurrences of query terms

We analyze highly ranked candidates for each query term and occurrences of query terms. Figure 2 shows the precision rates from the top candidate to among 10 candidates (the average for 30 query terms). The precision rate was more than 80 % among 3 candidates and it was more than 60 % among 10 candidates. For 30 query terms, there were 177 spoken documents (lecture speech) and 653 relevant segments. Therefore, each lecture includes about 4 relevant segments on average.

The above-mentioned analysis illustrated the highly ranked candidates are reliable and the query terms appear several times in the same spoken document. We utilize the knowledge for the re-scoring process and make correct but low ranked candidates rank higher to improve the STD performance.

B. Re-scoring: Giving a high priority in highly ranked documents

When we think about the *i*-th candidate in a spoken document DOC_A. If DOC_A includes query terms, it includes several query terms, as mentioned in the previous section and the average distance until *(i-1)*-th candidates in DOC_A is supposed to be small. It is because some of the *(i-1)* candidates is relevant and showed small distances. We introduce this idea to the following re-scoring process.

Re-scoring is carried out in the order from the top candidate to lower ranked candidate according to the CDP score. Let $D(l, i)$ to be the distance (CDP socre) for the *l*-th spoken document and the *i*–th candidate segment. $D(l, 1)$ at $i = 1$ in the equation (1) denotes the minimum distance of the *l*–th spoken document for the top candidate. The equation (2) denotes the new distance $newD(l, i)$ is given by adding the *i*–th original distance of the *l*-th spoken document and the average of the new distances from the top candidate to *(i-1)*–th candidates, linearly. The coefficient α is a weighting factor ($0 < \alpha <= 1$).

$$newD(l,i) = D(l,i) \quad (i = 1) \tag{1}$$

$$newD(l,i) = (\alpha \times D(l,i)) + (1-\alpha)\frac{\sum_{t=1}^{i-1} newD(l,t)}{i-1} \quad (i \neq 1) \tag{2}$$

If the average distance until *(i-1)*-th candidates (the second term in the Euation (2)) of a spoken document DOC_A is smaller than $D(B, 1) \ldots D(B, i)$ of a spoken document DOC_B, it can be considered that highly ranked candidates in DOC_A match query terms and DOC_A includes the query terms several times. Therefore, $newD(A, i)$ is made small (better) using the average distance until *(i-1)*-th candidates. The *i*-th distance is re-scored by using the average distance until *(i-1)*-th candidates in the same document.

IV. Evaluation Experiments

This chapter describes the evaluation experiments. First, the data sets and experimental conditions used in the experiments are described. Second, experimental setup about α described in the previous section is explained. Results and discussions are described lastly.

A. Data Sets and Experimental Conditions

Half of speech data in CSJ (Corpus of Spontaneous Japanese [13]) are used for training subword acoustic models and subword language models. The training data amount to about 300 hours including 1265 presentation speeches (14 minutes a presentation speech on average). Speech data in SDPWS (Corpus of Spoken Document Processing Workshop [9]) are used for the evaluation experiments, which amount to about 28 hours including 104 presentation speeches (16 minutes a presentation speech on average). Query terms and their relevant segments were released by organizers of SpokenDoc-2 task in NTCIR-10 [9]. We use two test data sets for evaluation experiments. The details of the test data sets are shown in Table I. Test set 1 and 2 were used at dry run and formal run query for SpokenDoc-2 of NTCIR-10.

Subword acoustic models and subword language models were trained using the HTK (Hidden Markov Model Toolkit [14]) and Palmkit [15] software tools, respectively. The feature parameters as extracted with HTK are shown in Table II together with the conditions for extracting the parameters. Julius rev. 4.1.5.1 [16] was used for the decoder of subword speech recognition.

TABLE I
OUTLINE OF TEST DATA SETS FOR EVALUATION

Test set	Spoken documents	Number of Queries
1	Corpus of Spoken Document Processing Workshop including 104 presentation speeches	30
2		92

For an evaluation measurement, we used the MAP (mean average precision) that is often used in the field of STD. MAP is computed, as follows. AP (average precision) for a query is obtained from equation (3) by averaging the precisions at every correct occurrence. In equation (3), *C* and *R* are the total number of correct segments and the lowest rank of the last correct segment, respectively. Let δi to be *1* if the *i*-th candidate section of query *s* is correct and *0* otherwise. Therefore, equation (3) averages the precision when a correct section is presented. MAP is obtained from equation (4) as the average of AP for each query *s*, where *T* is the total number of queries.

$$AP(s) = \frac{1}{C}\sum_{i=1}^{R} \delta_i \times precision(s,i) \tag{3}$$

$$MAP = \frac{1}{T}\sum_{s=1}^{T} AP(s) \tag{4}$$

B. Experimental Setup

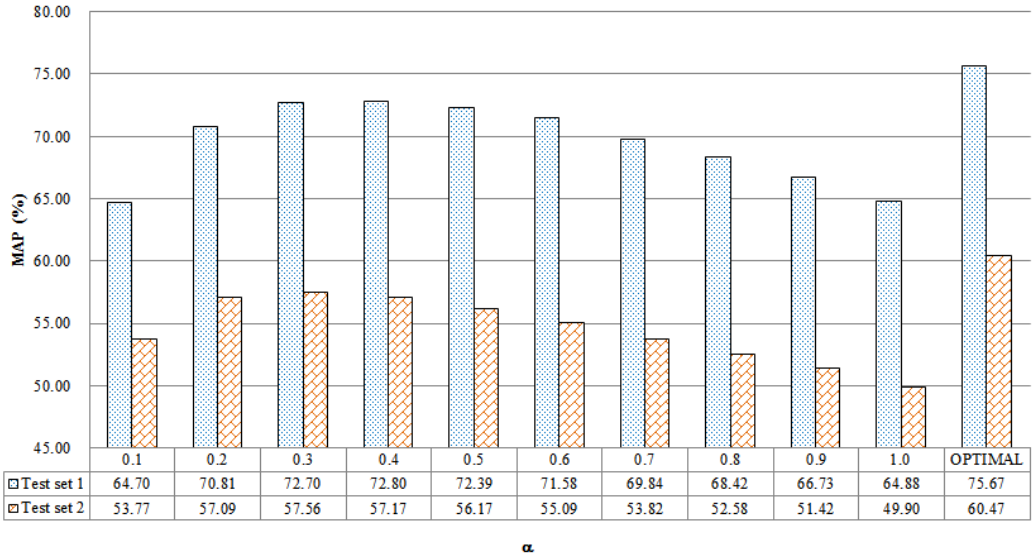
To evaluate the re-scoring method described in Section 3.B, we examine adequate value of the coefficient α in equation (2). We vary α from 0.1 to 1.0 by 0.1 for every query term. The reason why the case at α = 0.0 is excluded is that it is meaningless to ignore the actual distance of *i*-th candidate and the performance clearly degrades. In this paper we determine the coefficient α by a cross validation between the test sets 1 and 2.

C. Results at a constant α for all query terms

When the coefficient α is constant for all query terms and is varied from 0.1 to 1.0 by 0.1, the results are shown in Figure 3 for the two test sets. MAP at α = 1.0 denotes the performance when the proposed method is not applied. Therefore, the baseline performances without using the proposed method were 64.88% and 49.90% in MAP for the test set 1 and 2, respectively.

TABLE II
EXPERIMENTAL CONDITIONS

Sampling	16 KHz, 16 bits
Feature parameter	38 dim *MFCC + ΔMFCC + ΔΔMFCC + ΔPOWER + ΔΔPOWER*
Analysis window	Hamming window
Window length	25ms
Frame shift	10ms
Acoustic model	triphone Left-to-right HMM With 3 states
Language model	Syllable (251 models)



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	OPTIMAL
Test set 1	64.70	70.81	72.70	72.80	72.39	71.58	69.84	68.42	66.73	64.88	75.67
Test set 2	53.77	57.09	57.56	57.17	56.17	55.09	53.82	52.58	51.42	49.90	60.47

Fig. 3 Retrieval performance according α from 0.1 to 1.0.

The best performance was obtained at α=0.4 for the test set 1. At this α for the test set 2, MAP became 57.17% that was 7.27 points better than that (49.90%) without the proposed method. In the same way, the best performance was obtained at α=0.3 for the test set 2. At this α for the test set 1, MAP became 72.70% that was 7.82 points better than that (64.88%) without the proposed method.

These results demonstrated the effectiveness of the proposed method. OPTIMAL in the figure denotes MAP when α is controlled for each query so that the performance becomes best because α showing the best AP varies between query terms. We believe further performance improvement by adjusting α according to each query. This is a future work in the current paper.

V. Conclusions

Because highly ranked candidate is usually reliable and query terms are supposed to appear frequently in the target documents, we proposed a method giving a high priority to the candidate segments that are included in highly ranked documents to improve the retrieval performance in Spoken Term Detection. We conducted the performance evaluation experiments using open test collection for SpokenDoc-2 in NTCIR-10. Results showed the proposed method could improve more than 7.0 points in MAP, and demonstrated the effectiveness of the proposed method.

In the paper, the coefficient α is determined by a cross validation. We will control α for each query and determine it automatically for a future work.

Acknowledgment

This research is partially supported by Grand-in-Aid for Scientific Research (C) Project No. 24500124.

References

[1] C. Auzanne, JS. Garofolo, JG. Fiscus, and WM Fisher, "Automatic Language Model Adaptation for Spoken Document Retrieval," B1, 2000TREC-9 SDR Track, 2000.
[2] A. Fujii, and K. itou, "Evaluating Speech-Driven IR in the NTCIR-3Web Retrieval Task," Third NTCIR Workshop, 2003.
[3] P. Motlicek, F. Valente, and PN. Garner, "English Spoken Term Detection in Multilingual Recordings", INTERSPEECH 2010, pp.206-209, 2010.
[4] K. Iwata, *et al*., "Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity," INTERSPEECH, 2006.
[5] Roy Wallace, Robbie Vogt, and Sridha Sridharan, "A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation", INTERSPEECH 2007, pp2385-2388, 2007.
[6] N. Kanda, H. Sagawa, T. Sumiyoshi, and Y. Obuchi, "Open-Vocabulary Key word Detection from Super-Large Scale Speech Database", MMSP 2008, pp.939-944, 2008.
[7] Y. Itoh, *et al.*, "Two-stage vocabulary-free spoken document retrieval - subword identification and re-recognition of the identified sections", INTERSPEECH 2006, pp.1161-1164, 2006.
[8] C.-a. Chan, and L.-s. Lee, "Unsupervised Hidden Markov Modeling of Spoken Queries for Spoken Term Detection without Speech Recognition", INTERSPEECH 2011, pp.2141-2144, 2011.
[9] T. Akiba, *et al.*, "Overview of the NTCIR-10 SpokenDoc-2 Task", Proceedings of the NTCIR-10 Conference, 2013.
[10] H. Saito, *et al*., "An STD system for OOV query terms using various subword units", Proceedings of NTCIR-9 Workshop Meeting, pp.281-286, 2011.
[11] Y. Onodera, *et al.,* "Spoken Term Detection by Result Integration of Plural Subwords using Confidence Measure", WESPAC, 2009.
[12] F. Tanifuji, *et al*., "Improving perfomance of spoken term detection by appropriate distance between subwoed models", ASJvol2, pp.239-240, 2011-3.
[13] Corpus of Spontaneous Japanese, http://www.ninjal.ac.jp/csj/
[14] Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/
[15] palmkit, http://palmkit.sourceforge.net/
[16] Julius, http://julius.sourceforge.jp/