An STD system using multiple STD results and multiple rescoring method for NTCIR-12 SpokenQuery&Doc task

Ryota Konno
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g231n010@s.iwate-pu.ac.jp

Kazuki Ouchi
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g231n003@s.iwate-pu.ac.jp

Masato Obara
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g031k042@s.iwate-pu.ac.jp

Yoshino Shimizu
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g031k080@s.iwate-pu.ac.jp

Takashi Chiba
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g031k114@s.iwate-pu.ac.jp

Tatsuro Hirota
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
g031k140@s.iwate-pu.ac.jp

Yoshiaki Itoh
Iwate Prefectural University
Sugo 152-52 Takizawa,
Iwate, Japan
+81-19-694-2556
y-itoh@iwate-pu.ac.jp

ABSTRACT
Researches of Spoken Term Detection (STD) have been actively conducted in recent years. The task of STD is searching for a particular speech segment from a large amount of multimedia data that include audio or speech data. In NTCIR-12, a task containing multiple spoken queries is newly added to the STD task. In this paper, we explain an STD system that our team developed for the NTCIR-12 SpokenQuery & Doc task. We have already proposed the various methods to improve the STD accuracy for out-of-vocabulary (OOV) query terms. Our method consists of four steps. First, multiple automatic speech recognizers (ASRs) are performed for spoken documents using triphone, syllables, demiphone and SPS and multiple speech recognition results are obtained. Retrieval results are obtained for each subword unit. Second, these retrieval results are integrated [1][2]. Third, we apply a rescoring method to improve the STD accuracy that contains highly ranked candidates [3]. Lastly, a rescoring method is applied to compare a query with spoken documents in more detail by using the posterior probability obtained from Deep Neural Network (DNN) [4]. We apply this method to only the top candidates to reduce the retrieval time [5]. For a spoken query, we use two rescoring methods. First method compares two posterior probability vectors of the spoken query and spoken documents. Second method utilizes the papers in proceedings. We apply these methods to the test collection of NTCIR-12 and show experimental results for these methods.

Team Name
IWAPU

Subtasks
Spoken Query and Spoken Document Retrieval

Keywords
NTCIR-12, spoken term detection, spoken query

1. INTRODUCTION
Researches on spoken document retrieval (SDR) and spoken term detection (STD) have been actively conducted in an effort to enable efficient searching of the vast quantities of audiovisual data according to the recent and rapid increase in the capacity of recording media such as hard disks. STD is a task of locating matches in spoken documents to a query consisting of one or more words. In a typical STD system, speech data included in multi-media data were recognized by an automatic speech recognizer (ASR) in advance, and the recognition results were stored in inverted indexes. When query terms are given to the system, the system outputs retrieval results based on the inverted indexes.

Query terms are often out-of-vocabulary (OOV) words, such as technical terms, geographical names, personal names, and neologisms. In a word based STD system, it is impossible to find OOV query terms. Even if query terms are in-vocabulary (IV) words, the results of speech recognition are not always correct. Therefore, the problem of OOV query terms is important in the STD system. To enable to search OOV query terms, a subword-based system is often used. A subword is a smaller unit than a word such as monophone, triphone and so on. Subword recognition is performed for all spoken documents and subword sequences of spoken documents are prepared beforehand. When query terms are given to the system, the system converts the query terms to a sequence of subwords and then searches for the query's subword sequence among subword sequences of spoken documents. Matching between a query subword sequence and

subword sequences of spoken documents is conducted by a continuous dynamic programming (CDP) algorithm that performs a DTW (dynamic time warping) algorithm continuously. Although an edit distance is typically used for a local distance between the subwords, we use an acoustical subword distance [6] that is obtained from statistics of Hidden Markov Model (HMM) for each subword.

According to the wide spread of smart phones, voice becomes a familiar input means as used in Siri and Google voice search. A voice input for query terms is also considered to be a natural and needed way. In NTCIR-11 [7], a task using spoken queries (SQ) is introduced in addition to text queries that have been used in NTCIR so far. The problem of spoken queries is that it is difficult to determine whether the query terms are IV words or OOV words. Therefore, both results of a word based STD system and a subword based STD system are inevitably used. In NTCIR-11, we constructed a word based STD system and multiple subword systems for spoken queries. For subwords, we used four types of subwords such as triphone, syllable, demiphone and Sub-Phonetic Segment (SPS). We integrated these five STD results followed by a rescoring method and a rescoring method sequentially. The rescoring method decrease the distances of all candidates in a particular spoken document that a highly ranked candidate appears. The rescoring method realizes detailed matching by using the posterior probability of DNN that is applied to only the top candidates of the retrieval results because of its computation time.

In NTCIR-12, multi-term queries of SpokenQuery & Doc task was added. Multi term queries contain one or more words in a query with an audio file. We basically used the same method as in NTCIR-11. This time, we did not use demiphone and SPS and newly introduced two following rescoring methods. The first rescoring method used posteriorgram that is a sequence of all posterior probabilities generated by DNN frame-wisely. This method is also applied to only top candidates of the retrieval results because of the computation time. The second rescoring method used proceedings papers for spoken documents that were academic presentation speech. We explain the proposed method in section 2 and show the results of the experiments in section 3.

2. PROPOSED METHODS

This section describes STD methods we used in NTCIR-12. The proposed method consists of multiple methods. Some methods have been proposed in NTCIR-11, and the others are proposed in NTCIR-12. The proposed methods can be classified into following five methods roughly.

1. Segmentation for query terms in a spoken query
2. Matching between a spoken query and spoken documents
3. Rescoring of retrieval results
4. Integration of retrieval results
5. Integration of multiple query terms

From section 2.1 to 2.5, we describe each method in detail.

2.1 Segmentation for query terms in a spoken query

When a spoken query is given to the system, system splits the spoken query into multiple query terms. A recognition result obtained by a Kaldi decoder is split into multiple queries that consist of a phone sequence. If we observe short pauses for more than 30 frames, we regard the section as a boundary between query terms. When the boundary is not found, the query is regarded as a single query. When the boundary is found one or more, the query is regarded as multiple queries.

2.2 Matching between a spoken query and spoken document

In this section, we describe our retrieval method. The system allows two types of input that are text queries and spoken queries. If a text query is given, the query is converted to a subword sequence automatically according to Japanese conversion rules followed by matching a query subword sequence and subword sequences of spoken documents by CDP. If a spoken query is given, the spoken query is recognized by a word-based ASR and a subword-based ASR. Both recognition results are converted to subword sequences, respectively. CDP is performed as well as the case of the text query. An edit distance is often used as a local distance of CDP. We have proposed a subword acoustic distance taking account of acoustic similarity between any two subwords [6]. Acoustic distances between subwords are obtained from the statistics of HMMs that constitute subword acoustic models.

Each HMM consists of three states. After expanding a query subword sequence to a state sequence and subword sequences of spoken documents to state sequences, CDP matching is performed at a state level that is more detailed matching than at a subword level [8]. We use this state level matching when we apply CDP.

2.3 Rescoring of Retrieval Results

We use following four rescoring methods.

1. Rescoring using highly ranked candidates
2. Rescoring by DNN
3. Rescoring by posteriorgram
4. Rescoring using proceedings paper

2.3.1 Rescoring using Highly Ranked Candidates

We use a rescoring method to improve the retrieval accuracy after extracting candidate sections that are ranked by CDP distances [3]. We give a high priority to candidate sections contained in highly ranked documents by adjusting their CDP distances. The basic idea behind the proposed method is that highly ranked candidates are usually reliable and that a user selects query terms that are specific to and appear frequently in the target documents. Therefore, we prioritize the distances of candidate sections that appear in the documents that already contain highly ranked candidates according to the following equation,

$$D'(\Omega_j, k) = \alpha D(\Omega_j, k) + (1-\alpha)\frac{1}{T}\sum_{t=1}^{T} D(\Omega_j, t) \qquad (1)$$

where D and D'represents the CDP distance and the distance after rescoring, respectively. Ω_j and k are the utterance in the $j-$th document and the rank in Ω_j, respectively. The parameter α and T denote a weighting factor, and the number of candidates for rescoring. We call this rescoring method "High rank rescoring" in the paper.

2.3.2 Rescoring by DNN

We described a rescoring method using DNN. In 2.1, matching between a query subword sequence and subword sequences of

spoken documents is conducted by a CDP algorithm that performs the DTW algorithm continuously.

The proposed method performs detailed matching at a state level using more sophisticated local distances generated by DNN. The acoustic distances are more sophisticated than edit distances. In case that different two audio signals in spoken documents are symbolized into the same subword by an ASR, the acoustic distance between the symbolized subwords and a subword in a query become the same in spite of the difference between the two audio signals. The different posterior probabilities of the two audio signals are obtained by using a DNN. We introduce the probabilities output by a DNN to calculate local distances in CDP. An STD accuracy using the DNN at a state level is expected to be higher than that obtained by using conventional acoustic distances. To reduce the processing time, we apply the proposed method to only top candidate utterances that are ranked by our conventional STD method described in 2.1. Here, processing time largely depends on the number of candidate utterances, *K*. This method is so called "DNN rescoring" in this paper.

2.3.3 Rescoring by Posteriorgram

In this section, we described a rescoring method using posteriorgram [9]. For a spoken query, it might be possible to compare the spoken query and spoken documents directly at an audio level, a high STD accuracy, however, cannot be expected because of speakers' difference. To obtain a high STD accuracy, we should compare two posterior probability sequences between a spoken query and spoken documents. Posteriorgram is a posterior vector sequence. Let a local distance be a dot product between the two posterior probability vectors. More detailed matching at a frame level can be conducted compared with the matching at a state level. Although the dot product between the posterior probability vectors can be calculated at a high speed using the Graphics Processing Unit (GPU), the calculation requires a huge processing time when applying the method to all utterances. Therefore, the method is applied to only top K utterances to reduce the processing time. This method is so called "Posteriorgram rescoring" in this paper.

2.3.4 Rescoring using Proceedings Paper

We describe a rescoring method using proceedings paper. The method utilize the proceedings paper of lecture speech. The detail of the method appear in the autumn ASJ meeting. This method is so called "Paper rescoring" in this paper.

2.4 Integration of Retrieval Results

We have already proposed integrating plural results obtained from plural subword models for improving the retrieval performance, and confirmed the proposed method improved it robustly [1][2]. This method integrates the plural results linearly. Each subword model m $(1 \le m \le M)$ generates the distance $D_m(i,j)$ between a query Q_i $(1 \le i \le I)$ and an utterance or speech section S_j $(1 \le j \le J)$ and. Here, M, I and J denote the number of subword models, the number of queries, and the number of utterances, respectively. To integrate the retrieval results from plural subword models, the modified distance $D_{new}(i,j)$, which is a new criteria, is obtained by integrating the distances $D_m(i,j)$, according to the following equation,

$$D_{new}(i,j) = \sum_{m=1}^{M} \alpha_m D_m(i,j) \qquad \left(\sum_{m=1}^{M} \alpha_m = 1\right) \qquad (2)$$

Table 1 Conditions for DNN

Number of nodes	Input layer: 1320 Hidden layer: 2048 Output layer: 3009
Number of hidden layers	5 layers

Table 2 Conditions of feature extraction

Sampling	16 kHz / 16 bit
Feature Parameter	40 dim. FBANK 40dim. ΔFBANK 40dim. ΔΔFBANK
Window length	25 ms.
Frame shift	10 ms.

Table 3 Parameters of rescoring methods

High rank rescoring	$\alpha = 0.5$ $T = 2$
DNN rescoring	$K = 1{,}000$
Posteriorgram rescoring	$K = 1{,}000$
Paper rescoring	$\alpha = 0.7$

Table 4 Parameters of integration

Language model of spoken document recognition		α_1	α_2
Retrieval result 1	Retrieval result 2		
WORD	WORD	0.5	0.5
WORD	SYLLABLE	0.7	0.3
SYLLABLE	SYLLABLE	0.5	0.5

Table 5 Parameters of integration for DNN rescoring and Posteriorgram rescoring

Rescoring method of Retrieval result 1	Language model of spoken document recognition in Retrieval result 2	α_1	α_2
DNN rescoring	WORD	0.3	0.7
DNN rescoring	SYLLABLE	0.3	0.7
Posterirogram rescoring	WORD	0.3	0.7
Posterirogram rescoring	SYLLABLE	0.3	0.7

where α_m is a weighting factor for the m-th subword model. All of the weighting factors α_m are given beforehand, and the distances are combined linearly according equation (**2**).

In this paper, we did not use the proposed method to integrate multiple subword STD results, but used it to integrate an original STD result and the result after applying the rescoring methods described above.

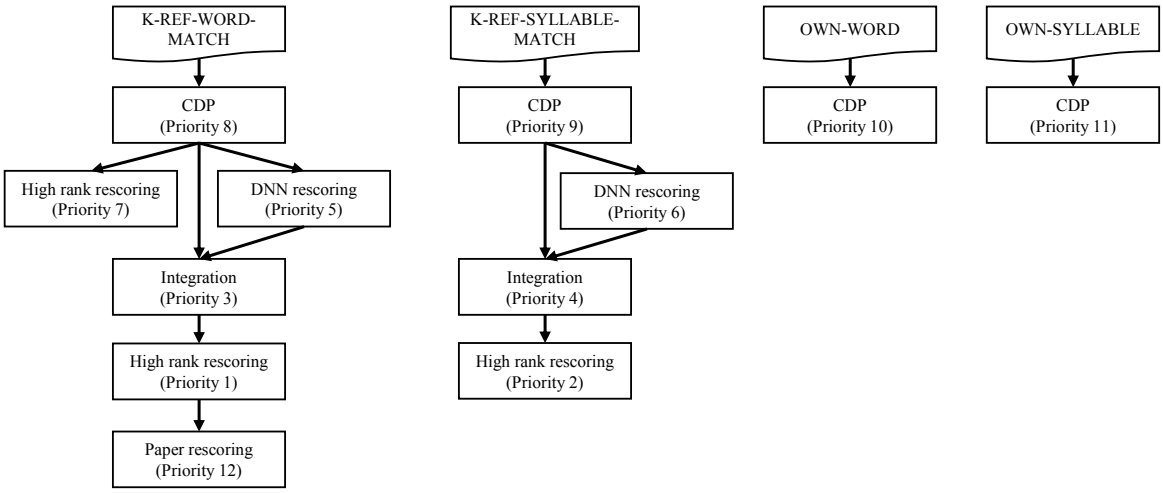

Fig. 1 Flowchart of submitted results for text query

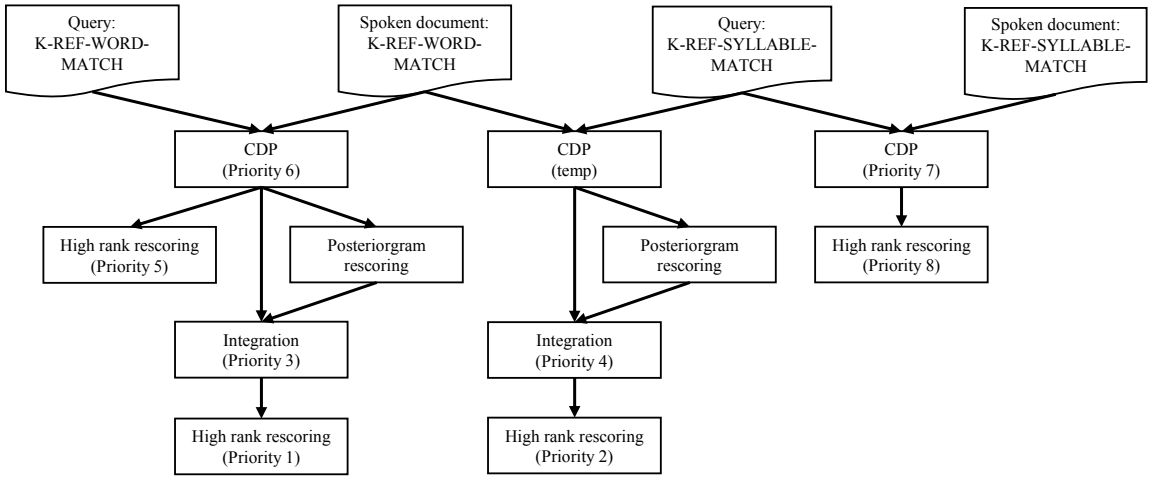

Fig. 2 Flowchart of submitted results for spoken query

2.5 Integration of multiple spoken query

We propose an integrating method for multiple spoken queries. In the case of multiple queries, the system splits a spoken query into multiple queries, and STD is performed for each query term, separately. The system integrates each STD result using the method described in 2.4.

3. EVALUATION EXPERIMENTS

3.1 Experimental Conditions

We used four recognition results. Two recognition results were provided by the organizer (K-REF-WORD-MATCH and K-REF-SYLLABLE-MATCH) and the other two recognition results were prepared by ourselves (OWN-WORD and OWN-SYLLABLE). We used a Julius ASR composed of DNN-HMM for OWN-WORD and OWN-SYLLABLE recognition. The DNN-HMM contains acoustic models with 3,009 states. The DNN was trained under the condition shown in Table 1. Feature vectors to input to the DNN were extracted under the conditions shown in Table 2. This DNN is also used in DNN rescoring and Posteriorgram rescoring. We used a word language model for OWN-WORD provided by the organizer and a syllable language model for OWN-SYLLABLE trained by even lectures of Corpus of Spontaneous Japanese (CSJ). We used parameters for rescoring methods and integration methods shown from Table 3 to Table 5. These parameters were optimized using the dry run test set in NTCIR-12.

3.2 Submit Data

We submitted 12 STD results in the text query task shown in Fig. 1, and 8 STD results in the spoken query task shown in Fig. 2.

In the text query task, we performed the retrieval method described in 2.2 against four types of recognition results (K-REF-WORD-MATCH, K-REF-SYLLABLE-MATCH, OWN-WORD and OWN-SYLLABLE). We gave a priority as follows.

- Priority from 8 to 11 was given to each of the four single STD results. These STD results are baseline of our system in the text query task.
- Priority 7 was given to the result applying the high rank rescoring method described in 2.3.1 to the STD result of

priority 8. This result showed the effect of the high rank rescoring.

- Priority 5 and 6 were given to the results applying DNN rescoring to the results of priority 8 and 9, respectively. These results showed the effect of DNN rescoring.
- Priority 4 was given to the result that was generated by integrating the STD results of priority 9 and 6 as shown in Fig. 1.
- Priority 3 was given to the result that was generated by integrating the STD results of priority 8 and 5 as shown in Fig. 1.

We have confirmed that it is possible to improve the STD accuracy by integrating these results in preliminary experiments.

- Priority 1 and 2 were given to the results by applying high rank rescoring to the STD results of priority 3 and 4, respectively.
- In case of the text query, we can apply paper rescoring described in 2.3.4. Priority 12 was given to the result applying the paper rescoring method.

The order applying the proposed methods was determined according to the results of the NTCIR-12 dry run.

In the spoken query task, we used the two types of recognition results for spoken queries and spoken documents (K-REF-WORD-MATCH, K-REF-SYLLABLE-MATCH). Therefore, four types of STD results were obtained. We gave a priority as follows.

- Priority 6 was given to the result applying the conventional retrieval method described in 2.2 using both K-REF-WORD-MATCH for spoken queries and spoken documents
- Priority 7 was given to the result applying the conventional retrieval method described in 2.2 using both K-REF-SYLLABLE-MATCH for spoken queries and spoken documents
- Priority 5 and 8 were given to the results applying the high rank rescoring to the results of priority 6 and 7, respectively. These results showed the effect of the high rank rescoring.
- Priority 3 and 4 were given to the results applying the posteriorgram rescoring to the result of priority 6 and the result of "temp", respectively, as shown in Fig.2.
- Priority 1 and 2 were given to the results applying the high rank rescoring to the results of priority 3 and 4, respectively.

The order applying the proposed methods and the combination of the integration were determined according to the results of the NTCIR-12 dry run.

4. CONCLUTIONS

We constructed an STD system using our proposed methods.

5. ACKNOWLEDGMENTS

This work was supported by JSPS (C), KAKENHI, Grant Number 15K00241.

6. REFERENCES

[1] Yoshiaki Itoh, et al, "An Integration Method of Retrieval Results using Multiple Subword Models for Vocabulary-free Spoken Document Retrieval," Proc. of INTERSPEECH 2007, pp.2389-2392, 2007.

[2] Yuji Onodera, et al, "Spoken Term Detection by Result Integration of Multiple Subwords using Confidence Measure," WESPAC, 2009.

[3] Kazuma Konno, et al, "Re-ranking of candidates using highly ranked candidates in Spoken Term Detection," ASJ, pp.191- 194, 2012-9.

[4] Geoffrey E. Hinton et al, "A Fast Learning Algorithm for Deep Belief Nets, Neural Computation," Vol. 18, pp. 1527-1554, 2006.

[5] Ryota Konno, et al., "Rescoring by a Deep Neural Network for Spoken Term Detection," APSIPA ASC 2015, pp.1207-1211, 2015.

[6] Kohei Iwata, et al., "Open-V ocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity," INTERSPEECH, 2006.

[7] Tomoyoshi Akiba, et al., "Overview of the NTCIR-11 SpokenQuery&Doc Task," Proc. of the 11th NTCIR Conference, pp.350-364, 2014.

[8] Naoki Yamamoto, et al., "Using Acoustic Dissimilarity Measures Based on State-level Distance Vector Representation for Improved Spoken Term Detection," APSIPA ASC 2013, 2013.

[9] Masato Obara, et al., "Improvement of STD search accuracy by combining DNN posteriorgram with a conventional method," ASJ, 2016.