An IWAPU STD System for OOV Query Terms and Spoken Queries

Jinki Takahashi
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231m022@s.iwate-pu.ac.jp

Takumi Hashimoto
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231m029@s.iwate-pu.ac.jp

Ryota Kon'no
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031j052@s.iwate-pu.ac.jp

Shota Sugawara
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031j076@s.iwate-pu.ac.jp

Kazuki Ouchi
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031j017@s.iwate-pu.ac.jp

Satoshi Oshima
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031j021@s.iwate-pu.ac.jp

Takahiro Akyu
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031j001@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

We have been proposing a Spoken Term Detection (STD) method for Out-Of-Vocabulary (OOV) query terms integrating various subword recognition results using monophone, triphone, demiphone, one third phone, and Sub-phonetic segment (SPS) models[1][2]. In this paper, we describe two methods for text OOV query terms and spoken queries. For text OOV query terms, we introduce four unique methods. First, we integrate multiple retrieval results obtained from multiple subword recognition[4][5]. Second, we use Deep Neural Network (DNN)[6] for computing output probabilities of Hidden Markov Models (HMM). Third, we apply a re-ranking method [7] utilizing highly ranked candidates. Fourth, DNN is also used for re-ranking for the retrieval results containing organizer's results. For spoken queries, we use speech recognition results of several speech recognizers including our word-based HMM and DNN-HMM recognizer, our syllable-based HMM and DNN-HMM recognizer and google voice. A few retrieval results obtained by each recognizer are combined. In STD tasks (SDPWS) of IR for Spoken Documents in NTCIR-11, we submit 15 types of retrieval results. For text query terms, we use transcriptions of our various speech recognizers, only an organizer's transcription, both of our transcriptions and an organizer's transcription, and so on. For spoken queries, we also use the same transcriptions of spoken documents, as mentioned above. We also submit a run using organizer's transcriptions for spoken documents, followed by our re-ranking methods.

Categories and Subject Descriptors

I.2.7 [**ARTIFICAL INTERLLIGENCE**]: Natural Language Processing – Speech recognition synthesis, *Text analysis*.

Team name

IWAPU-EX3

Subtasks

Spoken Term Detection (moderate-size task)

Keywords

IWAPU, Japanese, Spoken Term Detection, Spoken Query, Subword Model, SDPWS, DNN-HMM, HRC-re-ranking, DNN-re-scoring, Integration

1. INTRODUCTION

According to the rapid progress of information technology and the increase of the capacity of the recording mediums such as a hard disk or an optics disk in these years, every user comes to have much opportunity to deal with multimedia data such as video data that are available on such hard disk video recorders or the Internet. Recently, SDR (Spoken Document Retrieval) and STD (Spoken Term Detection) have been hot topics among speech processing researchers to deal with such enormous amount of data that are regarded as spoken documents[2][3][8][9]. In case of a common STD system, it generates a transcription of speech data using a large vocabulary continuous speech recognition (LVCSR) system, and finds query terms in the transcription. Although the method is advantageous in finding In-Vocabulary (IV) query terms at high speed, it has a difficulty in detecting Out-Of-Vocabulary (OOV) query terms that are not included in a dictionary of the LVCSR system, because OOV terms in spoken documents are inevitably

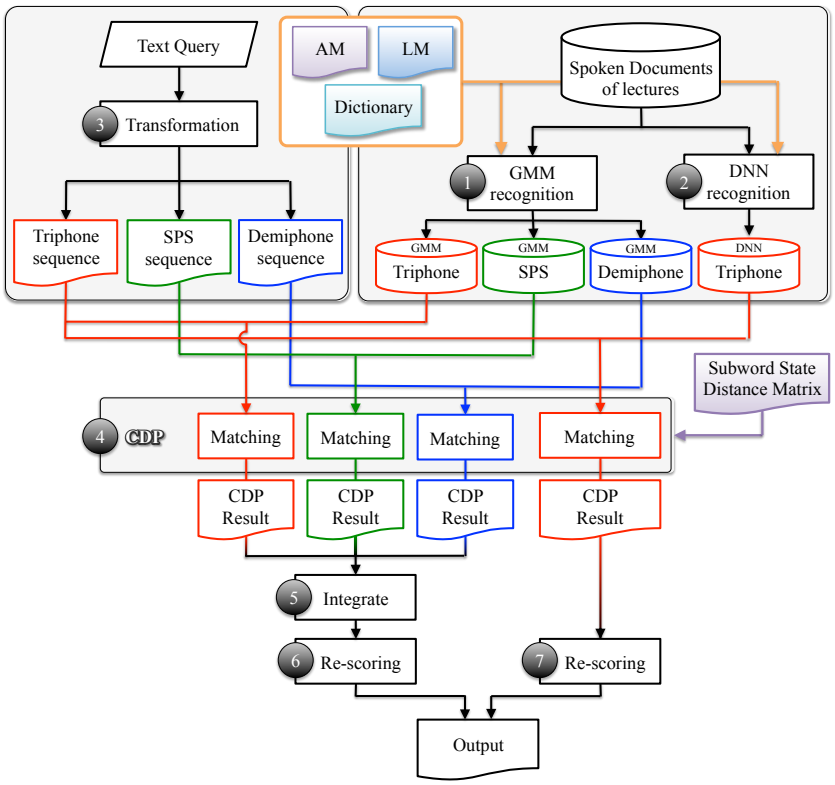

Figure 1: Outline of the STD method using multiple subword recognition results.

substituted to other words in the dictionary. STD systems must be able to detect OOV query terms because query terms are likely to be OOV terms, such as technical terms, geographical names, personal names and neologism and so on. To realize the detection of OOV query terms, a method using subword such as monophone and triphone is representative [1][10]. In NTCIR-10, we have proposed STD methods for OOV query terms using various subword units, such as monophone, triphone, demiphone, one third phone, and SPS models. For each subword model, the system compares a query subword sequence with all of the subword sequences in the Spoken Documents (SDs) and retrieves the target segments using Continuous Dynamic Programming (CDP) algorithm. Here, we introduce a phonetic distance between any two subword models for a local distance in CDP. In NTCIR-11, in addition to the integration method of multiple subword retrieval results [4][5] in NTCIR-10, we introduce three new methods to improve the STD accuracy and one four method to improve the retrieval time and index size.

- DNN-HMM recognition
- Re-ranking by highly ranked candidate
- Re-scoring by DNN
- Pre-retrieval by syllable bigram

Re-scoring was applied to improve the retrieval performance after CDP. We apply the most of the methods mentioned above to the STD tasks of IR for Spoken Documents in NTCIR-10 [2]. We use various subword models such as monophone, triphone, syllable, demiphone, and SPS. Phonetic distances between subword models are applied at a CDP process. Multiple STD results obtained from these subword models are integrated. Furthermore, we improve the performance by applying a longer N-gram language model. The performance is evaluated according to the criteria that the organizer provided.

The present paper describes the outline of our system first, and then our subword models, their acoustic models and language models are explained. Next, the integration method of multiple STD results is explained in detail after the explanation of subword based STD process using a single subword model and phonetic distances for a local distance of CDP and re-scoring. In Chapter 3, the performance of the proposed method is evaluated for the test collection of NTCIR-10. Lastly, conclusion is presented.

2. PROPOSED METHODS

2.1 System Configuration

The outline of the flowchart of our proposed system is illustrated in Figure 1. In our proposed system, we prepared subword acoustic models, a subword dictionary, subword language models, a syllable dictionary, and syllable language models, and a word dictionary, and word language models. SDs of lecture speeches are recognized beforehand using HMM and DNN-HMM. In the case of HMM, triphone/SPS/demiphone HMM acoustic models are used in subword recognition, syllable recognition, and word recognition (1). In the case of DNN-HMM, only triphone acoustic models are used in subword recognition, syllable recognition, and word recognition (2). When a text query is given to the system, it is converted to each subword sequence (3). For each recognition result, matching between a query subword sequence and subword sequences of spoken documents using CDP (4). Here, the acoustic distances between subword states are used for a local distance of CDP. Multiple CDP results are integrated in the same way as proposed in NTCIR-10 (5). Furthermore two rescoring methods are applied after the integration. One is re-ranking using highly ranked candidates (6), and another is DNN-HMM re-scoring (7). And output the result.

2.2 Recognition of STD

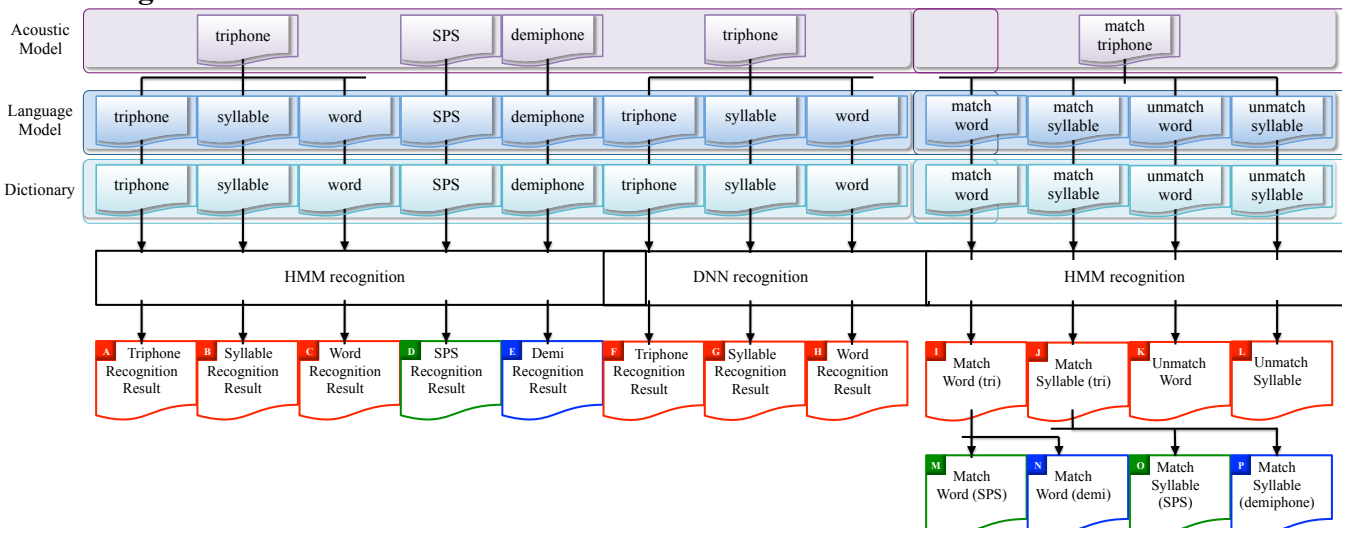

Figure 2: Using models and kind of recognition methods of SDs.

As shown in Figure 2, this section describes recognition of spoken documents in detail. We prepared triphone/SPS/demiphone acoustic models of HMM, and triphone acoustic models of DNN-HMM. Triphone/SPS/demiphone dictionaries and their language models are used for subword recognition. A syllable dictionary and syllable language models are used for syllable recognition. In the same way, a word dictionary and word language models are used for word recognitions. Therefore five types of recognition are conducted beforehand. In DNN-HMM, we prepare triphone/syllable/word dictionaries and their language models. Three types of recognition are conducted beforehand. Acoustic distance matrices between triphone states, SPS states, and demiphone states are prepared beforehand.

When a text query is given to the system, the text query is converted into triphone/SPS/demiphone sequences according to its monophone sequence automatically. Matching between the query and the recognition results **(A)～(J)** is performed by CDP. Local distances in CDP is obtained by only referring to an acoustic distance matrix.

2.3 Subword Models

This section describes subword models used in the paper. Three kinds of subword models, that is, monophone, triphone, sub-phonetic segment (SPS) and the demiphone [11][12] are used for subwords in the paper. These subword models and their sample descriptions of a monophone sequence "a k i" for each subword are shown in Table 1. Triphone is divided into two demiphone models: a model of the front part and a model of the rear part, as shown in Table 1. SPS models are designed so that they represent physical characteristics of pronunciation of consecutive phonemes. These subword models were confirmed to work well for STD [11].

Table 1: Subword models and "a k i" expressions.

Monophone	a		k		i		
Triphone	#-a+k		a-k+i		k-i+#		
SPS	#a	aa	ak	kk	ki	ii	i#
Demiphone	a1k		a2k		k1i	k2i	

2.4 Integration Method

We have already proposed integrating plural results obtained from plural subword models for improving the retrieval performance, and confirmed an improvement[4][5]. This method integrates the plural results linearly. Each subword model m ($1 \le m \le M$) generates the distance $D_m(i, j)$ between an utterance or speech section S_i ($1 \le i \le I$) and a query Q_i ($1 \le j \le J$). Here, M, I, and J denote the number of subword models, the number of utterances, and the number of queries, respectively. To integrate the retrieval results from plural subword models, these plural distances are simply combined lineally. This modified distance $D_m(i, j)$, which is a new criteria, is obtained by integrating the distances $D_m(i, j)$, according to the following equation,

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

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 to Eqs(1).

2.5 Re-ranking Method of Using Highly Ranked Candidates

We use a re-ranking method to improve the retrieval accuracy after extracting the candidate sections that are ranked by CDP distances[7]. 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 documents that already contain highly ranked candidates according to the following equation,

$$D'(\Omega_j,k) = \alpha \times D(\Omega_j,k) + (1-\alpha) \times \frac{1}{T}\sum_{t=1}^{T} D(\Omega_j,k) \quad (2)$$

where D and D' represents the CDP distance and the the distance after re-ranking, 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 re-scoring. We call this re-scoring method "HRC-re-ranking".

2.6 Using the DNN

2.6.1 DNN-HMM Recognition

We use the speech recognition method for estimating the output probability in DNN[6]. A DNN-HMM method is used for recognition.

2.6.2 DNN-re-scoring Method

We perform a subword based STD described in 2.1. The start and end time corresponding to the query terms is obtained according to subword recognition results. We perform CDP matching using DNN for only this interval to reduce the computation time of DNN. We call this re-scoring method "DNN-re-scoring".

2.7 Spoken Query Task

In the spoken query task, we use the results of "google speech" recognition and DNN word/syllable recognizers. Spoken queries are provided in multiple voice files for each query. We select one recognition result among various recognition results of multiple voice files using multiple recognizers. After deteniming phone sequence of the query, CDP matching is performed, followed by the integration of multiple retrieval results, and two re-scoring methods.

3. EVALUATION EXPERIMENTS

3.1 Training

Organizers distributed acoustic models, dictionaries and language models. A word dictionary and word language models trained by CSJ is so called "match word" **(I)**, and a syllable dictionary and syllable language models trained by CSJ is so called "match syllable" **(J)**. A word dictionary and word language model trained by newspaper articles are so called "unmatch word" **(K)**, and a syllable dictionary and syllable language model train by newspaper articles are so called "unmatch syllable" **(L)**. We convert **(I)** to SPS sequences **(M)**, and demiphone sequences **(N)**, convert **(J)** to SPS sequences **(O)**, and demiphone sequences **(P)**.

The conditions of feature extraction for acoustic models are listed in Table 2. The speech data of an actual presentation corpus of CSJ (Corpus of Spontaneous Japanese) are used for training data.

Table 2: Conditions of feature extraction for acoustic models.

Sampling	16kHz 16bit
Feature Parameter	12dim. MFCC+energy 12dim. ΔMFCC+Δenergy 12dim. ΔΔMFCC+ΔΔenergy
Window Length	25ms.
Frame Shift	10ms for monophone and triphone 5ms for demiphone and SPS

3.2 Condition

We used the parameters of the re-scoring method described in 2.5, when obtaining the best result in the previous NTCIR-10 (T = 2, α = 0.4). Also, parameter of the DNN-rescoring method described in 2.6. We used the parameters at a value of about 1 sec. processing time (K = 500) that is realistic retrieval time and was obtained from our previous results.

3.3 Submit Data

We submited 15 kinds of results for STD task, and 4 kinds of results for SQ task.

3.3.1 STD Task

1. Integrated **(C)** and **(H)** at a ratio of 1:1, and HRC-re-ranking
2. HRC-re-ranking **(H)**
3. HRC-re-ranking **(C)**
4. HRC-re-ranking **(G)**
5. DNN-re-scoring **(I)**
6. Integrated **(C)** and **(E)** at a ratio of 1:1, and HRC-re-ranking
7. DNN-re-scoring **(B)**
8. Integrated **(O)** and **(N)** at a ratio of 1:1, and HRC-re-ranking
9. Integrated **(P)** and **(N)** at a ratio of 1:1, and HRC-re-ranking
10. **(C)**
11. HRC-re-scoring **(N)**
12. DNN-re-scoring **(J)**
13. **(B)**
14. **(E)**
15. **(D)**

3.3.2 SQ Task

In SQ task, we submitted 4 results. We use "google speech," a DNN word recognizer and a DNN syllable recognizer to recognize spoken query files. For each query, we derive one phone sequence used in CDP from multiple recognizers results of spoken query files. We use 4 methods (a)～(d) for obtaining one phone sequence of multiple recognition results.

(a) Word recognition by DNN

(b) Syllable recognition by DNN

(c) The first candidate of the recognition result of "google speech" (When there is no candidates, the syllable recognition result by DNN is used)

(d) Selected recognition result of "google speech" (When there is no candidate, use the Syllable recognition result by DNN is used)

1. Use **(H)**, integrated (a), (b), (c), and (d) at a ratio of 40:10:5:1, and HRC-re-ranking
2. Use **(H)**, integrated (a), (b), and (c) at a ratio of 8:2:1, and HRC-re-ranking
3. Integrated **(C)**-(d) and **(H)**-(d) at a ratio of 1:1
4. HRC-re-ranking **(I)**-(d)

3.4 Necessary Resource

Table 3 shows the offline/online machine specification, the processing time, and the index size for 15 cases in STD and 4 cases in SQ tasks. Time in offline denotes the processing time for recognizing spoken documents, and time in online denotes the retrieval time including CDP, HRC-re-ranking/DNN-re-scoring and the integration of multiple results. Hyphen in the table shows machine specification is unknown.

Table 3: Machine Specification, Processing/Retrieval Time and Index Size.

Task / Submit No.		Offline			Online			Index Size [MB]
		Machine Specifications	Memory	Time [days]	Machine Specifications	Memory	Time [sec.]	
STD	1	Core i7-4770 3.9GHz	16GB	2.0	Core i7-4770 3.9GHz	16GB	0.58	2,150
	2						0.29	
	3	Penitum Xeon 2.56GHz(x2)	12GB	1.5	Penitum Xeon 2.56GHz(x2)	12GB	2.41	5,131
	4	Core i7-4770 3.9GHz	16GB	2.0	Core i7-4770 3.9GHz	16GB	2.46	
	5	-		-			2.42	
	6	Penitum Xeon 2.56GHz(x2)	12GB	8.5	Penitum Xeon 2.56GHz(x2)	12GB	1.16	2,155
	7			1.5			2.40	5,131
	8	-		-			1.10	1,610
	9	-		-			1.03	
	10	Penitum Xeon 2.56GHz(x2)	12GB	1.5			0.18	2,155
	11	-		-			0.52	1,610
	12	-		-	Core i7-4770 3.9GHz	16GB	2.40	5,131
	13	Xeon E3-1275v2 3.5GHz	16GB	1.5	Core i7-930 2.8GHz	12GB	0.28	1,761
	14			7.0			0.72	1,143
	15			3.0			0.78	728
SQ	1	Core i7-4770 3.9GHz	16GB	2.0	Core i7-4770 3.9GHz	16GB	0.88	2,411
	2	-		-	Penitum Xeon 2.56GHz(x2)	12GB	0.05	-
	3	Core i7-4770 3.9GHz	16GB	2.0	Core i7-4770 3.9GHz	16GB	0.29	2,145
	4	-		-	Penitum Xeon 2.56GHz(x2)	12GB	0.22	1,762

4. CONCLUSIONS

We constructed an STD system using our proposing methods that include the integration multiple STD results obtained using various subword units, two re-scoring methods, the introduction DNN to STD, and pre-retrieval by syllable bigrams. Experimental results showed the proposed method was able to achieve high STD accuracies compared with the baseline provided by the organizers in a realistic retrieval time (less than 1 sec.).

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Iwata K., Itoh Y., Kojima K., Ishigame M., Tanaka K. and Lee S., "Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity," INTERSPEECH, 2006.

[2] Tomoyoshi Akiba, et al., "Overview of the NTCIR-10 Spoken Doc-2 Task," Proceedings of the NTCIR-10 Conference, 2013.

[3] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, Gareth J. F. Jones, "Overview of the NTCIR-11 SpokenQuery&Doc Task," Proceedings of the NTCIR-11 Conference, Tokyo, Japan, 2013.

[4] 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.

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

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, Brian Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97 (2012).

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

[8] Auzanne C., Garofolo J. S., Fiscus J. G., Fisher W. M., "Automatic Language Model Adaptation for Spoken Document Retrieval," B1, 2000TREC-9 SDR Track, 2000.

[9] Petr Motlicek, Fabio Valente, Philip N, "Garner English Spoken Term Detection in Multilingual Recordings," INTERSPEECH 2010, pp.206-209, 2010

[10] Roy Wallace, et al, "A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation," INTERSPEECH 2007, pp.2385-2388, 2007.

[11] Iwata K, et al, "An Investigation of New Subword Models and Subword Phonetic Distance for Vocabulary-free Spoken Document Retrieval System," IPSJ Journal, Vol.48, No.5, pp. 1990-2000, 2007.

[12] Tanaka. K, et al, "Speech data retrieval system constructed on a universal phonetic code domain," ASRU'01 IEEE, pp.323-326, 2001.